UNIVERSIDADE FEDERAL DE SÃO CARLOS
CAMPUS DE SOROCABA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
PPGEP-S

APLICAÇÃO ESTRUTURADA DE DADOS DE REDES SOCIAIS NA MODELAGEM DE INSTRUMENTOS DE APOIO ÀS DECISÕES DE CONCESSÃO DE CRÉDITO

MARCOS FATTIBENE

Sorocaba
2015

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CAMPUS DE SOROCABA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO - PPGEPS

APLICAÇÃO ESTRUTURADA DE DADOS DE REDES SOCIAIS NA MODELAGEM DE INSTRUMENTOS DE APOIO ÀS DECISÕES DE CONCESSÃO DE CRÉDITO

MARCOS FATTIBENE

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção, do Campus Sorocaba, como parte dos requisitos à obtenção do título de Mestre em Engenharia de Produção.

Orientador: Prof. Dr. Jorge Luís Faria Meirelles

Sorocaba

2015

Fattibene, Marcos.
F254a Aplicação estruturada de dados de redes sociais na modelagem de instrumentos de apoio às decisões de concessão de crédito. / Marcos Fattibene. – – 2015.
69 f. : 28 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, *Campus* Sorocaba, Sorocaba, 2015
Orientador: Jorge Luís Faria Meirelles
Banca examinadora: Ana Elisa Périco, Flávio Leonel de Carvalho
Bibliografia

1. Análise de crédito. 2. Redes sociais on-line. 3. Modelos lineares (Estatística). I. Título. II. Sorocaba-Universidade Federal de São Carlos.

CDD 332.7

Ficha catalográfica elaborada pela Biblioteca *Campus* Sorocaba.

MARCOS FATTIBENE

"APLICAÇÃO ESTRUTURADA DE DADOS DE REDES SOCIAIS NA MODELAGEM DE INSTRUMENTOS DE APOIO ÀS DECISÕES DE CONCESSÃO DE CRÉDITO"

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção do Centro de Ciências e Tecnologias para a Sustentabilidade da Universidade Federal de São Carlos para obtenção do título de mestre em Engenharia de Produção, Área de Concentração: Gestão de Operações.
Sorocaba, 27 de janeiro de 2015

Orientador (a):

Prof. (a). Dr. (a). Jorge Luis Faria Meirelles
DEPS/UFSCar

Examinadores (as):

Prof. (a). Dr. (a). Ana Elisa Périco
UNESP-Araraquara

Prof. (a). Dr. (a). Flávio Leonel de Carvalho
DAdm/UFSCar

AGRADECIMENTOS

A D'US pelo dom da vida.

Aos meus pais por investirem seu amor, tempo e recursos. Por me instruírem nos sagrados valores e por me ensinarem a valorizar o aprendizado e o ensino como instrumento de alcance à cidadania.

Ao meu orientador, Prof. Dr. Jorge Luis Faria Meirelles, pelo seu apoio e acolhimento de um tema pioneiro, a despeito do risco envolvido e ao mentoreado dedicado e presente do Prof. Ricardo Rosene.

Às minhas amadas esposa, Carla, e filha Daniela, bem como aos meus irmãos José Paulo e Luiz Alberto pelo apoio, carinho e incentivo.

Ao Prof Dr. Ricardo Coser Mergulhão, meu primeiro contato na UFSCar-S: sua recepção e postura me incentivaram a desenvolver a Dissertação no PPGEP-S.

Aos Prof.s Dr.s José Geraldo Vidal Vieira e João Eduardo Azevedo Ramos da Silva pela dedicação ao ensino e pela oportunidade de aprendizado.

Ao PPGEP-S pela oportunidade de aprender e conviver com professores, funcionários e colegas e pelo possibilidade de desenvolver um tema inédito.

À minha tia Yvonne Fattibene, que sempre me incentivou à leitura e à escolha por cursar engenharia.

A titular da Secretaria do PPGEP-S, Érica Kushihara Akim, pelo inestimável apoio e exemplo pelo incomum idealismo e profissionalismo.

Aos professores da EAESP-FGV pelo apoio.

A todos os colegas e professores da pós-graduação em Engenharia de Produção, , com menção especial a Adriano de Moura Braatz, que tão cedo nos deixou.

À Escola Politécnica da USP, em especial ao Prof. Oswaldo Fadigas Fontes Torres, que me apresentou e influenciou a optar pela Engenharia de Produção.

Dedico esta dissertação,

Aos meus pais, Milton Fattibene e Elvira Santinha Sabatasso e à minha tia avó Alzira Oliveira do Carmo (in memorian). Sem eles, este momento crucial da minha trajetória não seria possível.

RESUMO

A análise de crédito para pessoas físicas tem tradicionalmente se apoiado em três pilares: comprovação documental de renda e de residência; consulta a birôs negativos de crédito, como SERASA Experian e SCPC e a utilização de modelos de projeção baseados na hipótese que perfis semelhantes reproduzirão no futuro o comportamento de crédito do passado, como por exemplo, os “credit scores” (HAND ; HENLEY, 2007). Tal abordagem tem se mostrado adequada, sendo, entretanto suscetível a momentos de crise econômica ou mudança rápida do perfil do mercado alvo, a exemplo do ocorrido no mercado imobiliário dos EUA no ano de 2008. O presente trabalho propõe-se indicar alternativas para a utilização do teor informacional presente nas Redes Sociais, onde os indivíduos registram suas opiniões, preferências e especialmente evidenciam sua rede de relacionamentos, no contexto da análise de risco de crédito. Evidenciaram-se formas de averiguação da premissa que proximidade de um indivíduo a outros com perfil de bons pagadores, ou vice-versa, influencia a taxa de adimplência. Para se ilustrar tais sugestões, foi utilizada uma rede social real, enriquecida com dados de crédito obtidos por simulação estatística. Foram elaborados três modelos de ponderação de dados e três modelos baseados em regressão linear múltipla. Em geral os resultados não foram estatisticamente significantes, dada a necessidade de uso de rede social estrangeira como também da geração de dados sintéticos de “score” de birô de crédito, dada a indisponibilidade de informações reais no País. Porém, ficou evidenciada a viabilidade da averiguação da hipótese de que o conteúdo informacional contido em redes sociais pode ampliar a eficiência do sistema de análise de crédito, se incorporado aos sistemas decisórios, operativos e de controle.

Palavras chave: análise de redes sociais, análise de crédito, modelagem estatística, regressão linear múltipla

ABSTRACT

The credit analysis for individuals has traditionally relied on three pillars: documentary proof of income and residence; refers to negative credit bureaus as SERASA and SCPC and the use of forecasting models based on the hypothesis that similar profiles in the future will reproduce the same credit behavior of the past, such as the "credit scores" (HAND; HENLEY, 2007) . This approach has been adequate, while being susceptible to moments of economic crisis or to fast profile changing of the target market, as occurred in the U.S. subprime in 2008. This study aims to point out ways to use Social Networks informational content, where individuals express and record their opinions, preferences, and especially get evident their network of relationships, in the credit analysis context. It was made evident the feasibility to investigate the assumption that an individual's proximity to other appropriate profile payers, or vice versa, influences the repayment rate. To illustrate such a conclusion, a real social network, enriched with credit data obtained by statistical simulation, was used. Three models of data weighting and three other based on multiple linear regression models were developed. In general the results were not statistically significant, by need to use a non-brazilian social network, as well synthetic data bureau score, since real information was not available in this country. It was shown a way to investigate the hypothesis that the informational content of a social network may generate greater efficiency into credit analysis when added to decision-making, operational and control systems of this segment.

Key words: social networks analysis, credit analysis, statistical modeling, multiple linear regression

LISTA DE TABELAS

LISTA DE QUADROS E FIGURAS

LISTA DE SIGLAS E ABREVIATURAS

ARS - Análise de Redes Sociais

BACEN - Banco Central do Brasil

Big Data – instrumental para captura, armazenagem e captura de volumes astronômicos de dados

EUA - Estados Unidos da América

FEBRABAN - Federação Brasileira de Bancos

GE - General Eletric

GM - General Motors

IBOPE - Instituto Brasileiro de Opinião e Estatística

p.p. - Pontos Percentuais

PIB - Produto Interno Bruto

PPGEP-S - Programa de Pós-Graduação em Engenharia de Produção - Campus Sorocaba

SERASA - Centralização de Bancos S/A

UFSCar - Universidade Federal de São Carlos

1. INTRODUÇÃO

1.1. OBJETIVOS

Objetivo geral

O objetivo geral deste trabalho é ampliar os conhecimentos sobre fontes alternativas de dados para suportar os sistemas e processos de decisão de crédito, com o propósito de aprimorar a eficácia dos processos de apoio a decisão de crédito. Novos conjuntos de dados, especialmente os que resguardam em grande parte a integridade das informações, caso das Redes Sociais, devem propiciar maior efetividade nos modelos em que são utilizados, especialmente se agregados aos dados e informações já tratados.

Objetivos específicos

Os objetivos específicos deste trabalho são:

- Revelar a viabilidade da inferência do grau de risco de crédito de determinado indivíduo - sem histórico suficiente de crédito - baseado na composição de pontuações de crédito dos indivíduos que apresentem algum grau de relacionamento com este, evidenciado por rede(s) social(is);
- Definir, para efeito de futura comparação, três modelos de ponderação de pontuação de crédito:
  i. O primeiro utilizou-se do conceito de “recência” do relacionamento entre dois atores em um contexto social, conceito introduzido por este trabalho para evidenciar os tempos de relacionamento entre indivíduos de determinada rede social. A hipótese subjacente é que quanto maior for a recência” entre dois indivíduos, maior seria a afinidade (homofilia) entre os indivíduos e maior também seria a correlação entre os seus comportamentos de crédito. O tempo é tradicionalmente uma variável considerada influente em modelagem de crédito, por exemplo, tempo no último emprego, tempo residindo na mesma residência;

ii. O segundo modelo incluiu o conceito de redes sociais que será apresentado adiante, a “centralidade de grau”, ou simplesmente “grau”, que mede o prestígio de um indivíduo em determinado grupo;

iii. O terceiro modelo se valerá da definição de “distância geodésica”, que também será comentada a seguir neste trabalho, e que fundamentalmente mede a proximidade entre dois indivíduos. A premissa é de que, quanto mais achegados forem dois atores, maior será a influência mútua no comportamento de crédito;

- Demonstrar que será possível confrontar o desempenho dos modelos que utilizaram o conjunto de dados extraídos da rede social. Foram aplicados critérios objetivos de desempenho via teste estatístico não paramétrico.

1.2 JUSTIFICATIVA

Segundo o BACEN, em maio de 2014 o volume total de crédito correspondeu a 56,1% do PIB, ante 56% em abril e 54,5% em maio do ano anterior, R$ 2.804 bilhões. Neste mesmo mês, a inadimplência do sistema financeiro, referente a operações com atrasos superiores a noventa dias, considerando-se as operações com recursos livres e direcionados, situou-se em 3,1%, com aumento de 0,1 p.p. no mês e redução de 0,5 p.p. em doze meses. O montante inadimplente foi de R$ 86,9 bilhões. Sendo assim, dado o impacto da indústria financeira na economia, faz-se necessário aprimorar os instrumentos de avaliação de crédito, enquanto a instituição do birô positivo evolui em abrangência, dada sua recente regulamentação.

A relevância das redes sociais cresce rapidamente no Brasil. Segundo dados do Instituto Brasileiro de Opinião Pública e Estatística – NIELSEN IBOPE (2014) havia na época 120,3 milhões de brasileiros com acesso à internet em algum local. Deste total, 90,8% está conectado às redes sociais. Nas redes sociais as pessoas registram seus dados cadastrais, expressam suas opiniões e mais ainda, se conectam e se desconectam a outros indivíduos. Trata-se de um verdadeiro laboratório do comportamento humano contemporâneo. Tal contexto se torna favorável para que a indústria de crédito e outros setores usem uma fonte alternativa de dados e informações para subsidiar suas operações.

2. PANORAMA DO CRÉDITO AO CONSUMIDOR NO BRASIL

Como Thomas, Edelman e Crook (2002) registraram, o crédito ao consumo tem cerca de 3000 anos de história, desde o tempo dos babilônios. Durante os últimos 750 anos, a indústria do crédito aos consumidores tem evoluído sobremaneira, começando com os usuários da Idade Média, mas o mercado de massa de consumidores no mundo não-islâmico é um fenômeno dos últimos sessenta anos. Na década de 1920, Henry Ford e Sloan haviam reconhecido que não era suficiente produzir produtos, como carros, para o mercado de massa, mas que também deveriam ser desenvolvidas formas de financiar sua compra. Isto conduziu ao desenvolvimento de empresas financeiras, como por exemplo, a GE Capital e a GM Finanças. O advento de cartões de crédito, na década de 1960, permitiu aos consumidores financiar boa parte de suas compras, como itens de consumo cotidiano, vestuário, viagens de férias, entre outros.

Em relação ao uso de ferramental de apoio à Análise de Crédito em massa no exterior e no Brasil, até o início do sec XX a avaliação era julgamental, e dependia da expertise e experiência de cada analista de crédito. Em 1928 e novamente em 1932, Paul FitzPatrick decidiu efetuar a comparação entre indicadores financeiros de empresas inadimplentes e não inadimplentes. Mais adiante, Altman, Haldeman e Narayanan (1977) se propuseram em elaborar modelo preditivo, com o uso da variável “z” padrão, utilizando-se do contraste entre indicadores de desempenho empresas -53 inadimplentes vs 58 adimplentes-, já utilizando técnicas estatísticas. Desde o início do século passado pouco mudou na essência da avaliação de crédito, em especial no Brasil. O tripé de decisão é o Score, a comprovação documental de renda e residência e a pontuação do birô negativo de crédito, com o uso de informações declaradas pelo proponente ao crédito o nas bases de dados internas das empresas.

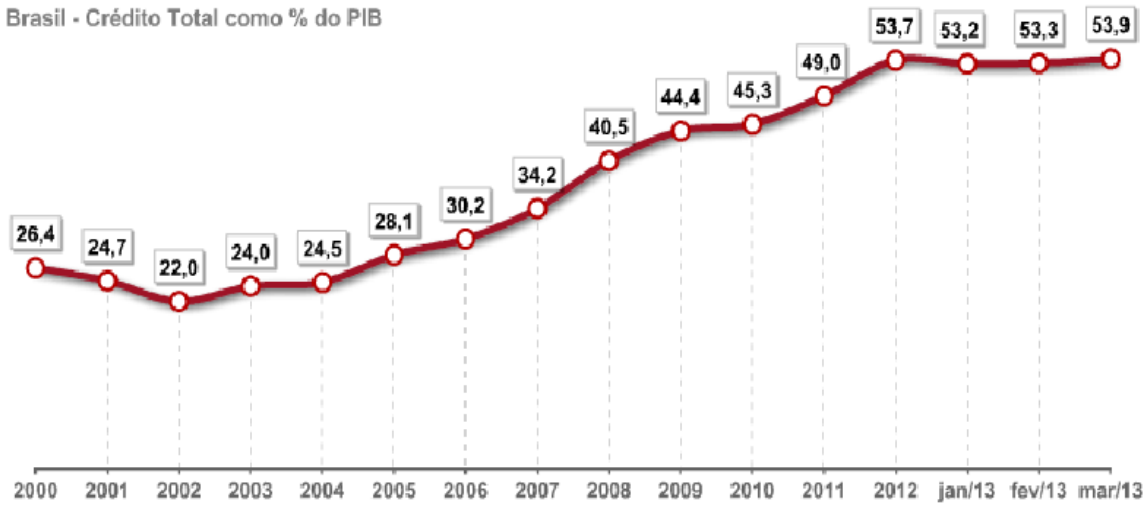
Anos 80, EUA, dissemina-se o uso de scores para a padronização de decisões e ganhos de produtividade, focos à época. Em meados dos anos 80, o Brasil começa a utilizar esta tecnologia, com uso de dados declarados e informações internas a cada instituição. Introdução do “Behavioral Score”, que utiliza dados internos das empresas que concedem crédito para prever

comportamento dos consumidores. Sucedeu-se um “boom” na concessão de crédito pessoa física em massa, com a ampliação da diversidade de produtos de financiamento e dos volumes concedidos e da massa de consumidores envolvidos. Há inclusive impactos econômicos e sociais sobre aqueles que têm suas solicitações de crédito negadas.

Dado que os cartões de crédito e débito têm sido muitas vezes utilizados em substituição aos cheques e ao dinheiro em espécie, tem havido nas últimas décadas uma significativa transformação nos mecanismos de pagamento. A maioria da população adulta tem algum produto financeiro a partir de um banco ou outra instituição financeira, e muitos têm mais do que um produto. Grandes bancos normalmente têm milhões de clientes e realizam bilhões de transações por ano. A enormidade do papel da dívida do consumidor de varejo é sugerida pelo fato de que a dívida média de um brasileiro é de cerca de cinquenta centavos de real, a cada real de rendimento disponível, segundo o Relatório Anual da FEDERAÇÂO BRASILEIRA DE BANCOS (FEBRABAN) (2013).

O crescimento do crédito aos consumidores nos últimos dez anos é considerável, conforme demonstrado na figura 1. O volume de crédito que era de cerca de 22% do PIB em 2002, ultrapassou os 50% do PIB, em 2012.

Figura 1: Volume de Crédito concedido no Brasil de 2000 a 2013

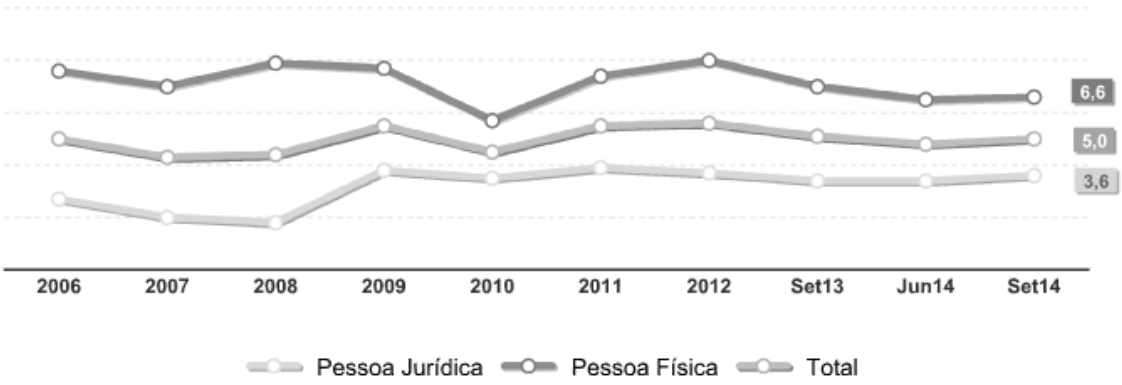


Fonte: Banco Central do Brasil (Abril/2013)

As principais ferramentas para a estimativa de risco de crédito utilizadas no País são o Credit e o Behavior Scoring (PRADO et. al., 2003), o primeiro

baseado em informações declaradas na proposta de obtenção de crédito e, a segunda, utilizada para os já clientes das instituições concedentes, incorporando informações sobre o comportamento de pagamento presente nas bases de dados internas destas instituições. Apesar do nível de endividamento médio do brasileiro se apresentar estável a partir de 2010, conforme mostra a figura 2, o volume total de inadimplência beira 3% do PIB, o que merece atenção das autoridades monetárias. Em específico quanto ao crédito para financiamento de veículos e na modalidade consignada, o Banco Central viu-se obrigado a tomar medidas restritivas, em Dezembro/2010, dados os sinais de uma possível bolha de crédito nestes dois canais de escoamento de recursos.

Figura 2: Patamares de inadimplência no Brasil de 2006 a 2013



Fonte: Banco Central do Brasil (Abril/2013).

Dado o contexto atual, em que se buscam instrumentos alternativos de avaliação de crédito por conta de fatores como mudança acelerada no perfil psicossocial dos brasileiros, a ascensão das classes emergentes e mesmo o aprendizado dos indivíduos, de quais os tipos de informação declarada são computados positivamente à concessão do crédito almejado, no momento de sua aplicação, novas abordagens são requeridas para aumentar o poder de inferência das instituições que concedem empréstimos e serviços financeiros. Cabe lembrar o aspecto da assimetria de informações entre o indivíduo que solicita um empréstimo (que tem toda a informação sobre seu histórico e intenções) e a organização emissora (que tem parte da informação sobre o histórico e intenções do proponente).

O contexto apresentado evidencia a necessidade de aprimoramento do instrumental de avaliação de risco para ampliar a eficiência empresarial dos sistemas financeiros e, em casos mais extremos, viabilizar econômica e financeiramente a atividade de concessão de crédito, melhor ilustrado pela tabela 1.

Tabela 1: Ilustração do impacto almejado na introdução de dados de redes sociais na avaliação de risco de crédito

Análise de Crédito	**Situação Atual**	**Situação Proposta**
Uso de dados de Redes Sociais	Esparso, as consultas e capturas de dados são manuais	Intensivo, as consultas e capturas de dados são automatizadas
Utilização de informações comportamentais aliadas às cadastrais	Parcial via consulta a birôs negativos de crédito (dados sobre restrições), se intensificando à medida que dissemine a utilização de birô positivo (dados sobre endividamento e desempenho de crédito)	Pleno, com o uso de dados de birôs negativos e positivos de crédito aliados às informações de redes sociais, seja na elaboração dos modelos preditivos ou em matrizes n-dimensionais de decisão
Efetividade da avaliação de risco	Parcial, dada a queda de desempenho dos modelos de discriminação ao longo do tempo, agravada quando ocorre mudança apreciável do perfil da população de proponentes de crédito	Crescente (por expectativa), com a incorporação de dados comportamentais recentes e fidedignos aos modelos de discriminação, tornando praticamente irrelevante o eventual efeito negativo de mudanças no perfil da população de aspirantes a crédito

Fonte: Elaboração do autor

As redes sociais sempre existiram na sociedade humana, em associações, clubes, irmandades e escolas, entre outros. Com a difusão e a popularização da web, inegavelmente estas tem ocupado espaço crescente na sociedade moderna (BAKSHY et. al., 2012). As pessoas expressam publicamente seus hábitos, opiniões, diálogos com indivíduos conhecidos ou não e aspectos pessoais, como fotos, endereços eletrônicos de sua

preferência, entre outros. Este conteúdo informacional, entretanto, ainda está subaproveitado pelas instituições concessoras de crédito. Há iniciativas esparsas na área de cobrança de dívidas, investigação de fraudes e na análise de proponentes de alta renda, mas nenhuma das iniciativas, pelo menos as divulgadas na mídia, revelam um aproveitamento amplo e sistemático deste arsenal de informações. A análise de redes sociais (ARS), formalizada inicialmente por Scott (1999) tem evoluído de forma acelerada em escopo e difusão, permitindo que aspectos relevantes das relações sociais possam ser capturados, processados e aproveitados para apoio a decisões das organizações.

3. AS REDES SOCIAIS E SEU POTENCIAL DE USO

3.1 Tipologia das redes sociais

Algumas classificações acerca de redes sociais, apresentadas a seguir, são relevantes para este trabalho, dadas as configurações de seus atores e como fonte de balizamento para a verificação da representatividade da comunidade escolhida para a realização do trabalho.

3.1.1. Classificação de Redes Sociais Quanto à Topologia

A) Redes Randômicas: segundo Barabási (1999), o modelo de rede mais simples é o de rede randômica, definido por Erdös e Rènyi (1960). Segundo este modelo de rede mais simples, cada par de vértices tem a mesma probabilidade de conexão e essa conexão ocorre de forma independente das demais;

B) Redes de Mundo Pequeno (“Small Worlds”): Este fenômeno indica que uma rede possui uma distância pequena entre quaisquer dois vértices. A maior manifestação popular deste fenômeno é o conceito dos “seis graus de separação”, descoberto pelo psicologista social Stanley Milgram, em 1967. Em seu famoso experimento, Milgram concluiu que existe uma distância média de 6 graus entre os moradores dos EUA. O fenômeno do mundo pequeno parece caracterizar muitas das redes complexas. Watts e Strogatz, (1998), acrescentam que o fenômeno do “mundo pequeno” se baseia nos conceitos de “laços fracos” (weak ties) e “laços fortes” (strong ties). Mundos pequenos são formados por indivíduos com 'laços fortes', onde em geral todos se conhecem. Tornam-se sociedades “fechadas”, consistidas por círculos de amigos altamente conectados. “Laços fracos” conectam membros destes círculos fechados a outros indivíduos pertencentes a outros grupos, que exercem um papel crucial em nossa habilidade de comunicação fora do “mundo pequeno”;

X) Redes Livres de Escala (“Scale Free”): Segundo Barabási e Bonabeau (2003), uma rede livre de escala contém “hubs”, ou seja, nós com um

grande número de relacionamentos. Neste tipo de rede, a distribuição dos graus dos nós segue uma lei de potência, visto que a maioria dos nós tem poucas conexões e alguns poucos nós possuem uma grande quantidade de relacionamentos. As redes livres de escala apresentam grande robustez frente a falhas acidentais devido a sua topologia heterogênea. A remoção randômica de alguns nós incide principalmente nos nós menores, que existem em maior quantidade na rede. Entretanto, as redes livres de escala são extremamente vulneráveis a ataques coordenados, que podem desconectar os seus “hubs” e acabar comprometendo a comunicação em toda a rede. As redes livres de escala também possuem um forte caráter epidêmico. Como os “hubs” estão conectados com muitos outros nós, pelo menos um “hub” tende a ser contaminado por um nó menor. Depois que um “hub” for infectado, ele vai espalhar o vírus para vários outros nós, eventualmente comprometendo também outros “hubs”, até atingir todo o sistema. Então, todos os vírus, mesmo os menos contagiosos, irão se espalhar e persistir no sistema. Entender os pontos fortes e fracos da robustez da rede pode ajudar a proteger sistemas vulneráveis a ataques ou sabotagem, a planejar campanhas de vacinação ou de marketing e até mesmo a monitorar e evitar um efeito cascata no caso de falências no sistema financeiro. Apesar de muitas redes sociais também serem redes livres de escala, existem exceções proeminentes. Dois exemplos são os sistemas de rodovias e energia elétrica dos Estados Unidos. Nas redes livres de escala também ocorre a conexão preferencial (BARABÁSI e BONABEAU, 2003). Quando novos nós aparecem, eles tendem a se conectar aos nós com mais conexões, e este mais popular passa a ter mais relacionamentos ao longo do tempo. Este processo também denominado de "rich gets richer", favorece os nós mais antigos da rede, que têm mais chance de virem a se tornar “hubs”.

3.1.2. Tipos de Redes Sociais

A. Quanto ao Escopo, na classificação de Hanneman e Riddle (2005):

i. Totais (também conhecidas como sócio-cêntricas): possuem um conjunto total de relacionamentos em uma unidade de análise (projeto, família, departamento, etc.);

ii. Egocêntricas: a maioria dos nós está conectada a nós simples ou individuais;

iii. Sistemas abertos: redes em que as fronteiras não são necessariamente claras.

B. Quanto aos Atores, segundo Wasserman e Faust (1994):

- One-mode: representam o relacionamento entre entidades sociais do mesmo tipo. Exemplo: quem é amigo de quem, quem pede conselho para quem, quem depende de quem;
- Two-mode: representam relacionamentos entre entidades sociais diferentes. Exemplo: as pessoas que foram a uma reunião, os desenvolvedores que corrigiram um determinando bug, pessoas que compõem organizações (Hanneman, Riddle, 2005). É importante ressaltar que a partir de uma rede two-mode podem-se obter as redes one-mode associadas a esta por meio de operações matemáticas na rede.

C. Quanto aos relacionamentos, de acordo com Wasserman e Faust (1994):

a. Díades: relação entre dois atores, é o nível mais baixo de rede;
b. Tríades: subgrupo de três atores e possíveis laços entre eles;
c. Grupo: Coleção de díades, tríades e subgrupos com um número finito de atores

3.1.3. Análise de Redes Sociais (ARS)

Segundo Wasserman e Faust (1994), redes sociais referem-se às relações formais e informais de um conjunto de pessoas (ou organizações ou outras entidades sociais), impulsionados por amizade, relações de trabalho ou compartilhamento de informações. Essas ligações são guias para a evolução desta estrutura social.

Dixom (2000) analisou o compartilhamento das informações, para discernir a posição e as ligações que os atores (participantes da Rede) mantêm nesta estrutura, buscando identificar o seu grau de influência. Partindo da premissa que o compartilhamento de informações e de conhecimentos entre as pessoas é constante, dados que estas apreciam compartilhar o que sabem. Ainda segundo a autora, as pessoas se sentiriam valorizadas quando há interesse de conhecer a sua expertise.

De acordo com Yu, Yan e Cheng (2001), cada ator teria informação abundante sobre sua situação, mas não sobre outras situações e pessoas. Para reduzir o grau de incerteza e consolidar a parceria, os atores anseiam por ter mais informações confiáveis sobre e de seus parceiros. Assim, todos lucrariam, porque cada ator construiria os alicerces e desenvolveria ações como base nas informações compartilhadas.

Dois indicadores da Análise de Redes Sociais em específico poderão ter papel importante na modelagem de crédito baseada em Redes Sociais: a centralidade e as ligações fortes e fracas.

- Centralidade: identificação dos atores que ocupam posições mais destacadas na Rede.
- Ligações fortes e fracas: as conexões fortes – as que contam com relacionamentos mais próximos-, e as fracas – mais distantes-, são analisadas tendo como base os índices de centralidade de proximidade, que será apresentado com mais detalhes a seguir.

i. Centralidade

De acordo com Freeman (1979), Centralidade é um recurso sociológico que não tem uma definição clara; é definido apenas de forma indireta. Um indivíduo é central em uma rede quando pode comunicar-se diretamente com muitos outros, ou está próximo de muitos atores, ou há muitos atores que o

utilizam como intermediário em suas comunicações. Uma interessante generalização deste conceito foi proposta por Opsahl, Agneessens e Skvoretz (2010) para redes em que as ligações entre os atores contam com pesos, sendo ponderado tanto o número de nós intermediários, quanto os pesos atribuídos aos vínculos. No presente trabalho, uma das formas utilizadas foi a ponderação com base na intensidade do relacionamento entre os atores.

Há cinco medidas principais para a centralidade:

i.1 Centralidade de Informação (“information centrality”);

Baseada no conceito de fluxo de informações, analisa todos os caminhos possíveis entre os atores. Por exemplo, tomemos alguns atores fictícios: se João se relaciona com José, José com Maria, Maria com Clara, e esta por sua vez, também está conectada com José, são possíveis dois percursos entre João e Maria:

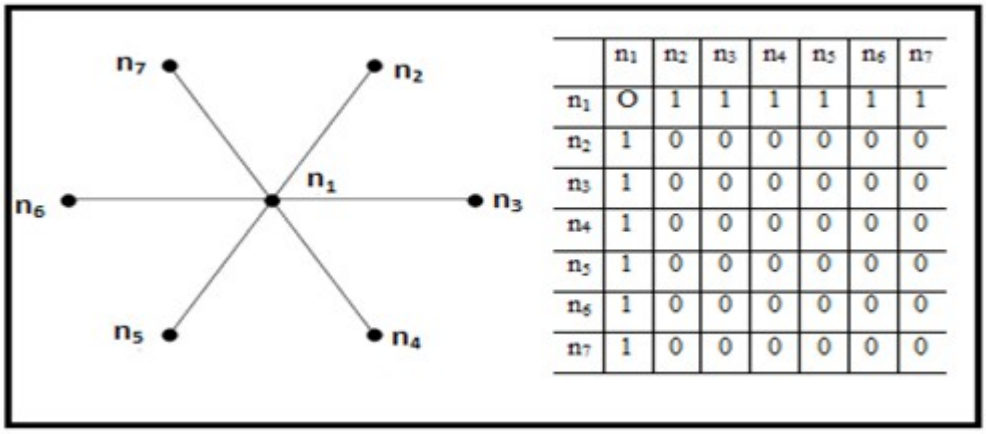
1. João- José-Maria, um percurso de tamanho 2;
2. João-José-Clara-Maria, este com tamanho 3.

Assim, para cada percurso analisado se faz possível apurar o fluxo de informação contida no caminho correspondente.

i.2 Centralidade de Grau (“degree centrality”);

Identifica o número de contatos diretos, ou seja, uma conexão que não depende de atores intermediários, que um ator mantém em uma rede: mede o nível de comunicação de um ator. Se este recebe muita informação – ligações direcionadas a ele – apresentaria destaque social ou teria prestígio nesta rede, ou seja, muitos outros atores buscariam compartilhar informações com ele. Importante lembrar que a influência e o prestígio são em parte responsáveis pelo fenômeno em que atores que, inicialmente, estão centralmente localizados em uma rede, tendem a se tornar ainda mais centrais (BARABÁSI e ALBERT, 1999), situação típica de uma “Power Law distribution”. A figura 3 ilustra o conceito: o ator n_1 possui grau 6.

Figura 3: Ilustração da Centralidade de Grau



	n_1	n_2	n_3	n_4	n_5	n_6	n_7
n_1	0	1	1	1	1	1	1
n_2	1	0	0	0	0	0	0
n_3	1	0	0	0	0	0	0
n_4	1	0	0	0	0	0	0
n_5	1	0	0	0	0	0	0
n_6	1	0	0	0	0	0	0
n_7	1	0	0	0	0	0	0

Fonte: http://dci.ccsa.ufpb.br/sic/?p=268 – visualizado em 29/Junho/2014

i.3 Centralidade de Intermediação (“betweenness centrality”);

Considera a possibilidade de um ator ser o meio para alcançar outros atores. Um indivíduo pode ter poucos contatos diretos na rede, mas exercer um importante papel intermediando informações. Como mediador ele tem o poder de controlar as informações que circulam na rede e o trajeto que elas percorrem.

i.4 Centralidade de fluxo (“flow centrality”)

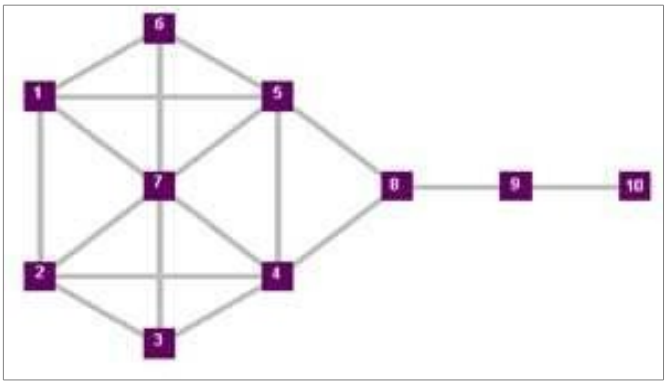
A medida de centralidade de fluxo analisa todos os caminhos possíveis para o contato entre os atores, complementando a medida de centralidade de intermediação, que analisa apenas o menor percurso, como apresentado no tópico i.1, também denominado caminho geodésico entre os atores. Aqui a intermediação mede-se pelo volume de fluxo entre os atores, que passa pelos caminhos em que está o ator. Alguns atores são mais centrais que outros no fluxo de dados e, portanto, têm maior controle das informações que fluem pela rede.

i.5 Centralidade de Proximidade (closeness centrality)

> Ressalta a distância de um ator em relação a outros, na rede. Este enfoque está baseado na distância geodésica – o menor número de passos em que dois atores podem ser ligados – de cada ator com todos os demais, considerando-se as distâncias tanto diretas quanto indiretas. Quanto maior a proximidade geodésica de um determinado ator com relação a outros atores da rede, mais central ele estará. Representaria independência, com a possibilidade de comunicação com muitos atores em uma rede, com um número mínimo de intermediários. A figura 4 ilustra o conceito supra: a maior distância geodésica ocorre entre os nós 1 (ou 2) e 10, que é de cinco; a menor é de um nó, entre 9 e 10, e vários outros pares de nós.

.

Figura 4: Ilustração da Centralidade de Proximidade



Fonte: Adaptado de: http://mande.co.uk/special-issues/network-models/ – visualizado em 29/Março/2013

Dentre as cinco medidas de centralidade, foi adotada a de “grau” para a aplicação ao presente trabalho, uma vez que sua apuração no conjunto de dados utilizado pôde ser mais precisa e verificável.

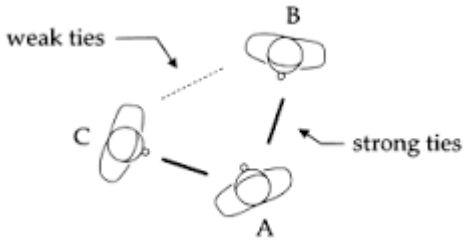
ii. Ligações Forte e Fracas

Segundo Levin, Cross e Abrams (2004), as pessoas que têm relacionamentos com maior distância geodésica (ligações fracas – “weak ties”) estão envolvidas em menor intensidade, enquanto que as com menor distância geodésica (ligações fortes – “strong ties”) têm um envolvimento maior. As ligações fracas são responsáveis pela baixa densidade em uma rede – em que muitas das possibilidades de relacionamento estão ausentes, enquanto que conjuntos consistentes dos mesmos indivíduos e seus parceiros mais próximos estão densamente ligados – muitas possibilidades de ligações estão presentes. Este é um dos principais conceitos de Redes Sociais em relação ao escopo deste trabalho.

Para ilustrar este conceito, os autores utilizaram os atores mais centrais da rede: denominados de “ego”. Os contatos diretos do “ego” são os “alters”: juntos eles formam uma rede “egocêntrica”. Os “alters” complementam o ego; pode-se considerá-los como a fonte de informação do “ego”; quanto mais bem relacionados forem os “alters”, mais informado será o “ego”, que usufrui do fluxo de informação capturado pelos atores que são seus “informantes”.

Além do conceito de centralidade de proximidade, outros atributos também são utilizados, entre os quais a duração (intensidade do contato) e a frequência (de raramente até mais de uma vez na semana). As ligações fracas, que o ego mantém, também apresentam certa relevância, pois acabam sendo pontes entre dois grupos de ligações fortes. Se um ator tem poucas ligações fracas pode ficar sem as informações que fluem em outros grupos densamente conectados. As ligações fortes também podem ser avaliadas por meio das “díades” – interação entre dois atores que trocam informações em que um considerou o outro como um dos seus contatos importantes na rede para o compartilhamento da informação, ilustrado na figura 5, a seguir.

Figura 5: Ilustração das ligações fortes e fracas em uma Rede Social



Fonte http://www.leadershipcloseup.com/2012/12/14/strength-of-weak-ties-in-social-networking-seek-to-be-worth-knowing/ – visualizado em 19/Outubro/2014

Tal aplicação de dados de redes sociais à análise de crédito passa por duas fases importantes: a primeira, da captura e organização de dados, inicialmente dispersos e desconexos. A segunda, passa pela utilização de um método que permita a estimação de riscos e propensões para sua aplicação na análise de crédito. A proposta do presente trabalho é incorporar à abordagem de “credit score” uma ponderação polinomial de fatores de influência no comportamento de crédito em determinado horizonte fixo de avaliação, os conceitos de homofilia - presunção de que o contato entre pessoas com características similares ocorre em maior frequência do que entre pessoas com baixa similaridade (MCPHERSON; SMITH-LOVIN e COOK, 2001) - e de heterofilia, em que características ou atributos diferentes funcionam como elemento atrativo nas relações. Os mesmos autores apontam que há uma ampla variedade de dimensões como raça, etnia, sexo, idade, religião, educação, ocupação, classe social, entre outras que pode levar à homofilia, assim como muitas outras, principalmente ligadas ao status, que levariam à heterofilia.

iii) Recência (tempo de relacionamento entre dois atores)

O tempo decorrido para determinadas relações é usualmente uma variável relevante para a estimação do comportamento de crédito. Tomemos o exemplo de “tempo de residência”, fator usualmente coletado em algumas propostas para obtenção de crédito, e que mais ainda, acaba fazendo parte dos “credit score” que participam como item importante para o processo de

seleção de proponentes.

Este trabalho propõe introduzir a variável “recência” - tempo de relacionamento entre dois atores - como uma candidata a explicar o grau de coesividade entre estes, e averiguar seu poder de discriminação em conjunto com a pontuação dos indivíduos no birô de crédito. A hipótese aqui é que se dois indivíduos apresentam alto grau de coesão, o comportamento de crédito, como outras dimensões de um relacionamento, apresentará correlação positiva.

3.2 O papel potencial das Redes Sociais na análise de crédito

Pesquisa preliminar sobre a utilização de dados estruturados capturados de Redes Sociais na análise de crédito demonstrou que a literatura sobre esse assunto ainda se faz embrionária. O trabalho que de alguma forma se alinha ao tema aqui proposto é o de Danyllo et al (2013), da área de computação, em que os autores tomaram uma lista de clientes de determinada instituição financeira, em que constava a informação da presença ou não destes em birô de crédito negativo. Posteriormente pesquisaram tais nomes no Twitter, encontrando 16% de usuários desta rede social. Tais nomes apresentavam maiores índices de restrição de crédito do que os demais clientes, que não se encontravam nesta rede social. Além disso, os usuários que não apresentavam restrições usualmente se conectavam a pessoas também sem restrições de crédito, em uma típica manifestação de homofilia. Song et. al (2010) discorreram sobre o potencial uso de redes sociais na avaliação e gestão de microcrédito. Em termos da previsão de comportamento, Fei et. al (2011) discorrem sobre um algoritmo que se propõe a prever a resposta que um determinado indivíduo dará a uma postagem na rede social que está utilizando no momento. Na área de crédito pode-se citar o trabalho de Wydick, Hayes e Hilliker (2011), em que se constata que pessoas que fazem parte de alguma rede social, aqui no sentido de comunidade paralela, como uma igreja, têm chances significativamente maiores de ter suas solicitações de crédito aprovadas.

Embora não se possa fazer uma correlação imediata entre a psique e o comportamento de crédito, cabe citar o trabalho de Kramer, Guillory e Hancock

(2014) que realizaram uma experiência massiva com cerca de 700 mil usuários do Facebook, onde constataram que os estados emocionais de certos indivíduos podem ser transferidos para outras pessoas através de contágio emocional, levando as pessoas a experimentar as mesmas emoções, sem que eles esteja conscientes disto. Eles forneceram evidências experimentais de que o contágio emocional ocorre sem a interação direta entre as pessoas - se expor a um amigo que expressa determinada emoção foi suficiente -, sendo que houve completa ausência de sinais não-verbais. Dessa forma, sendo o desempenho de crédito um comportamento psico-social, seria razoável crer que este deva ser influenciado pela rede de contatos, seja pela forma como foi constituída (laços de origem) e pela sua estrutura atual (relacionamentos de 1º, 2º, ... e Nº graus)

Um aspecto crítico para a explicação do comportamento de crédito de um indivíduo ou empresa é a reputação. Sabater e Sierra (2002), apontaram três dimensões da reputação: a do "indivíduo", que se refere às interações diretas entre dois agentes, a "social", que se refere ao grupo social a que um agente pertence (escopo das Redes Sociais), e a dimensão "ontológica", que acrescenta diferentes facetas da reputação.

A utilização de informações advindas de Redes Sociais na avaliação do risco de crédito de indivíduos aparenta ser promissora, dadas experiências bem sucedidas nos campos da seleção de pessoal[1], combate ao terror[2] e marketing & vendas, por exemplo.

---

1 ALMERI, T; MARTINS, K.; PAULA, D. O uso das redes sociais virtuais nos processos de recrutamento e seleção. Revista de Educação, Cultura e Comunicação Social da FATEA, Lorena, n. 8, v. 4, p. 77-94, jul./dez. 2013

2 KREBS, Valdis E. Mapping networks of terrorist cells. Connections, v. 24, n. 3, p. 43-52, 2002.

Entretanto, ainda há aspectos não muito claros sobre a viabilidade do uso de dados das redes sociais na análise de crédito. Da mesma forma que a ferramenta de scoring sofreu um processo de ajuste à legislação vigente nos EUA na década de 70, com o “Ato de Igualdade nas Oportunidades de Crédito” (CHANDLER; EWERT, 1976), o uso de dados de redes sociais ainda não tem suas fronteiras jurídicas claramente definidas no Brasil, devendo ser objeto de atenção dos juristas e dos empreendedores.

4. METODOLOGIA

O trabalho envolveu a seleção de uma rede social que representasse as comunidades de clientes típicos para a análise de crédito, a captura de seus dados para tratamento por softwares estatístico e de análise de redes sociais, a aplicação dos modelos matemáticos de ponderação e as análises de regressão linear múltipla e a averiguação da sua discriminação por testes estatísticos de hipóteses não paramétricos.

4.1. Seleção e Mapeamento da Rede Social para a realização do trabalho

A comunidade escolhida foi representativa de uma rede social com potencial de utilização pela indústria de crédito. Uma rede social "Livre de Escala" seria a ideal para a utilização neste trabalho, por ser típica das mais conhecidas como Facebook, Linkedin, Instragran, entre outras. Isto porque não há controle sobre o perfil de indivíduo que busca crédito, configurando-se em grupos heterogêneos. Entretanto, as bases de dados disponíveis tanto no Brasil quanto no exterior não incluem este tipo de rede. A alternativa seria capturar os dados de uma rede real, mas isto foi inviável em decorrência de dois principais obstáculos:

- inviabilidade operacional: tanto por parte da empresa detentora do software e da infra-estrutura das Redes Sociais eletrônicas, quanto dos indivíduos selecionados, que teriam de autorizar formalmente a utilização dos dados;

- limitação de orçamento: as empresas gestoras das redes sociais usualmente cobram para permitirem a utilização de suas bases e infra-estrutura e não havia recursos suficientes dedicados para tal.

Sendo assim, foi utilizada uma das bases de dados disponíveis no software de gestão de redes sociais utilizado neste trabalho, uma base de dados real com estrutura de "Mundo Pequeno". Ressalte-se que este tipo de configuração também se prestou aos propósitos do trabalho, já que no contexto de crédito é comum lidar com grupos homogêneos, por exemplo, quando se acessam grupos empresariais, associações, sindicatos ou cooperativas, em que o perfil dos integrantes apresenta pontos comuns.

Os dados da rede social adotada são reais, mas os nomes das pessoas foram omitidos pela Ucinet, para que não se considerasse haver invasão de privacidade e para evitar eventuais ações judiciais, já que ainda não há consenso sobre a legalidade da extração e armazenamento de dados presentes nas redes sociais eletrônicas sem a autorização expressa dos seus titulares. Trata-se de uma comunidade de 960 profissionais que trabalham em projetos de saúde no exterior. O critério adotado para a escolha foi a comunidade com maior número de atores possível. Os vínculos de 1º grau se estabelecem na medida em que os dois atores em questão estejam trabalhando ou tenham trabalhado em um mesmo projeto.

Foi considerada a utilização de uma comunidade de indivíduos no Brasil, mas o tempo e os custos da coleta dos dados seriam inviáveis, dados o prazo e o orçamento disponível para a realização do trabalho.

A seguir são apresentadas algumas características desta rede de relacionamentos a serem mapeadas:

- Centralidade média de Grau (ou Grau médio): número médio de indivíduos que se relacionam diretamente com cada ator;
- Densidade[1]: razão entre o número de pares de vínculos formados entre os atores pelo o número de pares de vínculos possíveis;
- Distância Geodésica média (ou distância média): número médio de intermediários necessários para colocar dois atores em contato;
- Desvio padrão da distância geodésica: desvio padrão dos passos necessários para conectar dois atores;
- Diâmetro[2]: 5 – a maior distância geodésica possível entre dois indivíduos;
- Tamanho da Matriz de Adjacência: número de atores ao quadrado.

A distribuição da centralidade de grau será confrontada com uma “Power Distribution”, cuja formulação foi primordialmente apresentada por Vilfredo Pareto, expressando concentração de renda e de prestígio – este último mais vinculado às características das Redes Sociais. Será elaborado um gráfico de duas dimensões, em que o percentual do eixo “y” representa o percentual agregado de indivíduos por faixa do eixo “x”, que abarca o grau dos

indivíduos em ordem decrescente. A adesão da curva real a uma “Power Distribution” será visual, dada a característica prospectiva do trabalho.

Os dados da rede PV960 serão exportados para o formato de “matriz de adjacência”, armazenados em planilha Excel. A matriz de adjacência lista os atores em linha e colunas e atribui o valor 1 quando houver vínculo direto entre dois atores e 0 quando isto não ocorrer. Trata-se, portanto de uma matriz simétrica. No caso específico da diagonal, as células foram deixadas em branco, pois para esta rede social não se aplicava haver valor diferente de zero.

4.2. Geração de dados sintéticos de birôs de crédito e de “recência” via simulação estatística

Esta etapa foi designada para se obterem valores para o pseudo score de birô. Como a identidade dos indivíduos foi mantida em segredo, não seria possível a coleta de valores reais para esta pontuação. Dessa forma, foi necessário gerar dados via simulação estatística.

Para tal, foram colhidas informações da distribuição de birô score de consumidores do Canadá (ALBERTA MORTGAGES, 2014)[3], produzido pela Equifax, empresa internacional de credit bureau, que foram ajustadas a uma distribuição Beta, com assimetria negativa, com $\alpha = 7$ e $\beta=2$. A amplitude usual é de 1000 pontos (planejada para oscilar entre 0 e 1000 pontos). Os dados de uma única empresa birô de crédito foram utilizados para averiguar a distribuição do score de birô almejada, pois não foram encontrados trabalhos científicos sobre o tema e tampouco dados de outras empresas que pudessem embasar levantamentos.

3 <http http://www.alberta-mortgages.com/articles/credit-bureau.html

Da mesma forma, os dados necessários para a apuração da variável “Recência” não estão disponíveis no banco de dados da rede PV960. Para gerar tal distribuição, foi adotada uma distribuição Weibull, adequada para modelar tempo de vida (BENDER, AUGUSTIN e BLETTNER, 2005), com parâmetro de escala de sete anos e de forma de dois, parâmetros estes que foram aleatoriamente definidos, também pela ausência de dados reais para embasar um levantamento.

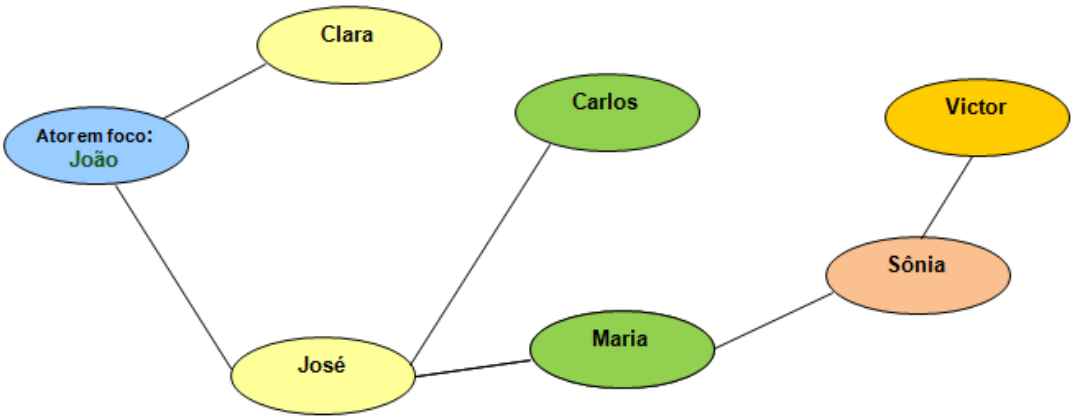
4.3. Os Modelos de Ponderação

O pressuposto básico do trabalho é que, para determinado ator em análise, o comportamento de crédito dos indivíduos que fazem parte de sua rede de relacionamentos influencia o comportamento do indivíduo em avaliação de maneira mais decisiva quanto maior for a proximidade entre estas pessoas conectadas ou grau de influência que estes atores apresentam. Dada a relativa simplicidade de aplicação e facilidade de replicação, foi adotada a ponderação linear simples.

Assim, foram aplicadas ponderações da pontuação de crédito dos contatos de determinado indivíduo com três variáveis: a distância geodésica, a centralidade de grau e a “recência”, como ilustrado na figura 6. Tais conceitos serão mais bem descritos nos tópicos hipotéticos e meramente ilustrativas a seguir:

- distância geodésica: em relação ao ator “João”, de uma rede fictícia. João e José são relacionamentos de 1º grau, assim a distância geodésica é de 1; entre Maria e João existe um intermediário, e a distância é 2; já entre Sônia e ele, 2 intermediários, com distância 3; já com Victor, são 4 passos, e portanto, a distância é 4;
- centralidade de grau: João conta com 2 contatos diretos, portanto seu “grau” é 2; José possui 3 contatos diretos, e assim, seu grau é 3;
- recência: João e José se conhecem há 2,5 anos, que é a “recência” entre eles.

Figura 6: Ilustração de indicadores de redes sociais



Fonte: Elaboração do autor

4.3.1 Ponderação da Pontuação de Crédito pela Distância Geodésica

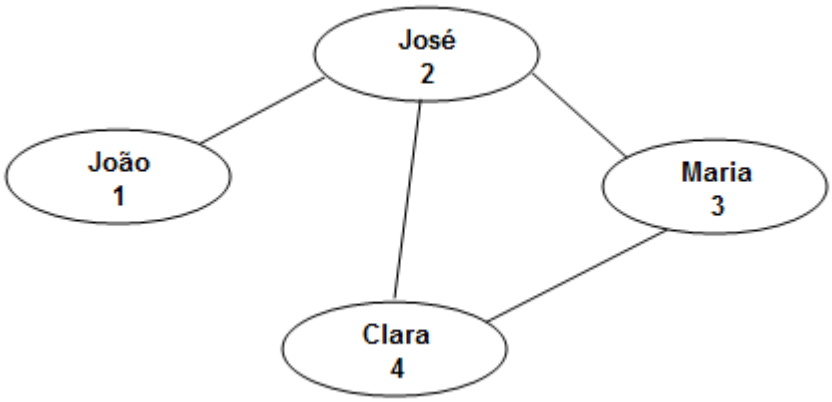
Seja uma rede de “N” indivíduos, e nela, um ator “i”, em que possui relacionamento com outros atores, representado pelo vetor $\mathbf{V}_i = (a_1, a_2, ..., a_i, ... ,a_{n-1}, a_n)$, sendo $a_j=1$, quando os atores “i” e “j” estiverem conectados, e $a_j=0$, quando estes atores não estiverem interligados. O índice “i” varia de 1 a N. Seja C_j, a centralidade de grau de indivíduo a_k, pertencente à rede em questão, com “k” variando também de 1 a N, e “j” variando de 1 a C_k, com i≠k. A título de exemplificação, considere-se a mesma rede utilizada no item 3.3.1.i.1, em que João se relaciona com José, José com Maria, Maria com Clara, e esta por sua vez também está conectada com José.

Assim, João é o indivíduo 1, José 2, Maria 3 e Clara 4. Então, N = 4 e os vetores representados para cada indivíduo e suas respectivas centralidades de grau, número de indivíduos com os quais se relacionam, são:

- $\mathbf{V}_1 = (0, 1, 0, 0)$, $G_1 = 1$;
- $\mathbf{V}_2 = (1, 0, 1, 1)$; $G_2 = 3$;
- $\mathbf{V}_3 = (0, 1, 0, 1)$; $G_3 = 2$;
- $\mathbf{V}_4 = (0, 1, 1, 0)$; $G_4 = 2$.

Note-se que a somatória de todos os elementos do vetor $\mathbf{V}_i$ é a centralidade de grau do indivíduo “i”.

Figura 7: Rede social fictícia utilizada para exemplificar os conceitos adotados



Fonte: Elaboração do autor

Conforme definição apresentada anteriormente, a distância geodésica será 1, quando dois atores estiverem conectados diretamente; 2, quando o ator "i" estiver conectado com outro indivíduo que conheça um terceiro ator; 3, quando o ator "i" estiver conectado com outro indivíduo que conheça alguém que conheça um quarto ator; e assim por diante.

Sejam também:

- **D** a matriz de distâncias geodésicas, $(d_{ij})_{NxN}$. Então, se o indivíduo "i" estiver diretamente conectado com o indivíduo "j", a distância geodésica será 1, ou $d_{ij} = 1$. Para atores que precisam de um intermediário para se conectar a outro indivíduo, a distância geodésica é de 2, no caso de 3 intermediários, a distância é de 3, e assim por diante. No exemplo:

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix}$$

- $\mathbf{P}_i$ o vetor com os valores da pontuação de crédito de cada indivíduo desta rede, com "i" variando de 1 até N. No exemplo, considere-se o segundo vetor:

- **P** = (600, 350, 800, 100).

Assim, a pontuação de João, no birô de crédito, é de 600, a de José, 350, a de Maria, 800 e a de Clara, 100 pontos

Dessa forma o valor de interesse é a pontuação média de birô de crédito de um indivíduo “h”, considerando a ponderação das pontuações dos indivíduos, $\mathbf{P}_k$ (k≠l) conectados pela distância $d_{l,k}$, representado pela linha “k” da matriz **D**. Para se calcular a média ponderada das pontuações, deve-se tomar o produto escalar entre os vetores $\mathbf{P}_k$ e a linha “k” da matriz $\mathbf{D}_{h,k}$, dividido pela somatória dos valores de $\mathbf{D}_{1l}$. Obtém-se assim o vetor $\mathbf{P}_h$, das pontuações de escore de crédito estimadas pela ponderação pela distância geodésica:

$$P_d = \frac{(P_k \,.\, \mathrm{D}_{\mathrm{h,k}})}{\sum_{k=1}^{N} D_{\mathrm{h,k}}}$$

Na rede social do exemplo:

P_1 = (0*600 + 1*350 + 2*800 + 2*100) / (0+1+2+2) = 390 ;
P_2 = (1*600 + 0*350 + 1*800 + 1*100) / (1+0+1+1) = 500 ;
P_3 = (2*600 + 1*350 + 0*800 + 1*100) / (2+1+0+1) = 412,5 ;
P_4 = (0*600 + 1*350 + 1*800 + 0*100) / (0+1+1+0) = 675 .

Assim, $\mathbf{P}_d$ = (390, 500, 412,5, 675), que é o vetor de estimativas da pontuação de crédito dos indivíduos utilizando-se como base a ponderação pela distância geodésica de seus relacionamentos.

Para o escopo deste trabalho, foi desprezada a influência de atores com mais de 3 níveis de distância, dada a palpável dificuldade computacional de apuração, hipótese que poderá ser testada em trabalhos futuros.

4.3.2 Ponderação da Pontuação de Crédito pela Centralidade de Grau

Seja uma rede de “N” indivíduos, e nela, um ator “i”, em que possui relacionamento de 1º grau com outros atores, representado pelo vetor $\mathbf{V}_i$ = (a_1, a_2, ...,a_i, ... ,a_{n-1}, a_n), com a_j=1, quando os atores “i” e “j” estiverem conectados, e a_j=0, quando estes atores não estiverem interligados. Sejam também:

- $\mathbf{V}_h$ os vetores com a indicação dos relacionamentos do indivíduo “h”. Se um outro ator se relaciona com ele, o valor de V_h. é 1, caso contrário, 0, com “h” variando de 1 até N.

- **P** o vetor com os valores da pontuação de crédito de cada indivíduo desta rede, com “h” variando de 1 até N.

O valor de interesse é a pontuação média de birô de crédito de um indivíduo “h”, considerando a ponderação das pontuações dos indivíduos, P_k (k≠h) conectados com ele pelo grau G_k. Para obter-se a média ponderada de P_h , deve-se tomar o produto escalar entre os vetores P_k e G_k, e dividindo-se pelo soma dos valores do vetor V_k – o grau propriamente dito-, com “k” variando de 1 até N.

Assim, P_h é calculado da seguinte forma:

$$P_g = \frac{(P_k . V_h)}{\sum_{k=1}^{k=N} V_k}$$

Na rede social do exemplo:

- $\mathbf{V}_1$ = (0, 1, 0, 0) ; G_1 = 1;
- $\mathbf{V}_2$ = (1, 0, 1, 1) ; G_2 = 3;
- $\mathbf{V}_3$ = (0, 1, 0, 1) ; G_3 = 2;
- $\mathbf{V}_4$ = (0, 1, 1, 0) ; G_4 = 2.

P_1 = (0*600 + 1*350 + 0*800 + 0*100) / (0+1+0+0) = 350 ;
P_2 = (1*600 + 0*350 + 1*800 + 1*100) / (1+0+1+1) = 500 ;
P_3 = (0*600 + 1*350 + 0*800 + 1*100) / (0+1+0+1) = 225 ;
P_4 = (0*600 + 1*350 + 1*800 + 0*100) / (0+1+1+0) = 675 .

Assim, $\mathbf{P}_g$ = (350, 500, 225, 675), que é o vetor de estimativas da pontuação de crédito dos indivíduos utilizando-se como base a ponderação pela grau de seus relacionamentos.

4.3.3 Ponderação da Pontuação de Crédito pela Recência (tempo de relacionamento entre dois atores)

Seja uma rede de “N” indivíduos, e nela, um ator “i”, em que possui relacionamento com outros atores, representado pelo vetor **V** = (a_1, a_2, ...,a_i, ... ,a_{n-1}, a_n), com a_j=1, quando os atores “i” e “j” estiverem conectados, e a_j=0, quando estes atores não estiverem interligados, com “j” variando de 1 a N.

Sejam também:

- $\mathbf{R}_h$ o vetor com os valores da Recência entre os relacionamentos de um par de atores, daqui em diante chamado de Recência de cada indivíduo desta rede, com “h” variando de 1 até N;
- **P** o vetor com os valores da pontuação de crédito de cada indivíduo desta rede, com “h” variando de 1 até N.

O valor de interesse é a pontuação média de birô de crédito de um indivíduo “l”, considerando a ponderação das pontuações dos indivíduos, P_k (k≠l) conectados com ele pela recência R_k. Para obtermos a média ponderada de P_h , devemos tomar o produto escalar entre os vetores P_k e R_k, diviindo-se o resultado pela soma das recências R_h, obtendo-se a média ponderada por **R**.

$$P_r = \frac{(P_h . \ R_h)}{\sum_{k=1}^{k=N} R_k}$$

Na rede social do exemplo:

- $\mathbf{R}_1$ = (0, 7,4 , 2,6 , 1,4) ;
- $\mathbf{R}_2$ = (7,4, 0, 2,3, 3,4) ;
- $\mathbf{R}_3$ = (2,6, 2,3, 0, 8,2) ;
- $\mathbf{R}_4$ = (1,4, 3,4, 8,2 , 0).

P_1 = (0*600 + 7,4*350 + 2,6*800 + 1,4*100) / (0+7,4+2,6+1,4) = 421,9 ;
P_2 = (7,4*600 + 0*350 + 2,3*800 + 3,4*100) / (7,4+0+2,3+3,4) = 505,3 ;
P_3 = (2,6*600 + 2,3*350 + 0*800 + 8,2*100) / (2,6+2,3+0+8,2) = 243,1 ;
P_4 = (1,4*600 + 3,4*350 + 8,2*800 + 0*100) / (1,4+3,4+8,2+0) = 660,8 .

Assim, $\mathbf{P}_r$ = (421,9, 505,3 , 243,1, 660,8), que é o vetor de estimativas da pontuação de crédito dos indivíduos utilizando-se como base a ponderação pela recência de seus relacionamentos.

4.4 Modelagem da influência do comportamento das conexões de determinado ator via regressão linear múltipla

Modelos multivariados envolvem análise do relacionamento entre múltiplas variáveis explicativas e, em alguns casos, múltiplas variáveis dependentes. Grande parte dos estudos efetuados para averiguar o efeito exercido por duas ou mais variáveis independentes sobre uma variável dependente utiliza a análise de Regressão Linear Múltipla. O presente trabalho utiliza tal ferramenta pela robustez, simplicidade e disseminação pela comunidade acadêmica e empresarial, A Regressão Linear Múltipla (RLM) é definida por Tabachnick (2001) como um conjunto de técnicas estatísticas que possibilita a avaliação do relacionamento de uma variável dependente com diversas variáveis independentes. Frequentemente as variáveis independentes correlacionadas entre si, e se não houver critérios consistentes para a exclusão de variáveis, pode se desconsiderar itens importantes na explicação da variável em foco. Nestes casos, é mais segura a utilização de técnicas estatísticas como a RLM. Embora esta técnica seja sensível à natureza correlacionada dos preditores, suas limitações já são bastante conhecidas.

O resultado final de uma RLM é uma equação de um hiperplano que representa a melhor predição de uma variável dependente a partir de diversas variáveis independentes. Esta equação representa um modelo aditivo, no qual as variáveis preditoras somam-se na explicação da variável critério. A equação da regressão linear pode ser representada por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k xi_k + \varepsilon_i \text{ , com } i = 1, \ldots, N \quad (1)$$

Onde:

- "N" é o número de atores da comunidade estudada;
- Y é a variável de interesse (variável dependente ou variável resposta), y_i a ocorrência amostral da pontuação do i-ésimo ator;
- As variáveis Xj (independentes ou covariáveis) e $X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ o vetor de observações das variáveis independentes para o i-ésimo indivíduo, , sendo $j = 1, 2, \ldots, k$, com $k < N$;
- $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)'$ o vetor de coeficientes de regressão (parâmetros);
- ε_i é o componente de erro aleatório. Assume-se que esses erros são independentes e seguem distribuição normal com média zero e variância desconhecida σ^2.

O modelo (1) é chamado de regressão linear múltipla, pois envolve mais de um coeficiente de regressão a se estimar. O termo "linear" indica que o modelo é linear em relação aos parâmetros $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$, e não porque y é uma função linear dos x'_s.

Para que o uso desta equação seja eficaz na predição da variável dependente em estudo, o pesquisador deve examinar previamente os pressupostos da RLM, bem como identificar as conseqüências da sua violação. Entre os pressupostos citados por Tabachnick (2001), estão: (1) a multicolinearidade, (2) a singularidade, (3) a homogeneidade nas variâncias, (4) a normalidade e (5) a linearidade. Dada a utilização preponderante de dados sintéticos no presente trabalho, foi adotada somente a verificação da normalidade das varáveis explicativas, expostas na seção de resultados. Além disso, embora seja imprescindível que o pesquisador examine esses pressupostos antes de iniciar suas análises, nota-se que a RLM é um modelo eficaz contra a violação de grande parte dos pressupostos. Por exemplo, no caso da inclusão de variáveis multicolineares ou singulares nas análises, o pesquisador estará perdendo graus de liberdade, o que consequentemente reduziria o poder estatístico de suas conclusões. O pesquisador pode, ainda, estar excluindo de seu modelo de estudo variáveis importantes para a explicação do fenômeno em questão, as quais podem estar correlacionadas com uma variável multicolinear. A violação do pres-

suposto de normalidade pode ser atenuada por meio do aumento do tamanho da amostra da população pesquisada (MARTINS e DOMINGUES, 2011).

Para a seleção das variáveis explicativas, pode-se definir modelos de maneiras diferentes, tanto do ponto de vista de quais variáveis comporão o modelo, quanto como elas se apresentarão no modelo. O ajuste de todos os possíveis modelos conta com diferentes metodologias de escolha a serem adotadas para a escolha dos modelos:

- A metodologia stepwise é bastante adequada para seleção automática de modelos. O método stepwise subdivide-se em três estratégias: forward, backward e both (passo-a-frente, passo-a-trás e passo-a-passo). Na estratégia forward, inicialmente ajusta-se um modelo sem variáveis e a cada passo, avaliam-se as variáveis candidatas a tomarem parte do modelo. Se houver variável com contribuição significativa de explicação para o modelo, esta é adicionada a ele. O procedimento repete-se até quando não houver variável que trará contribuição significativa. Na estratégia backward, inicialmente ajusta-se um modelo com todas as variáveis. A cada passo, avalia-se cada variável para a saída do modelo, e se houver variável que não traz contribuição significativa ao modelo, esta sai do modelo. O procedimento repete-se até que permaneçam apenas variáveis significativas para ele. Já a estratégia both, inicialmente ajusta-se um modelo sem variáveis e no primeiro passo, avalia-se se alguma das variáveis trará contribuição significativa ao modelo. Caso haja, esta variável é incluída no modelo. A seguir, a cada passo o grupo de variáveis que está no modelo é avaliado para sair do modelo e o grupo de variáveis que não estão no modelo é avaliado para entrar no modelo.

Para verificar a viabilidade de se obterem previsões mais elaboradas que as ponderações avaliadas na seção anterior foram testadas três configurações de regressão linear múltipla:

- a primeira utilizando as ponderações de centralidade de grau, ponderação da distância geodésica até o grau 2 e a ponderação pela recência;.
- a segunda, utilizando os valores brutos da centralidade de grau, a média da distância geodésica até o grau 2 dos atores vinculados ao ator em avaliação e a média da recência dos atores vinculados ao ator em avaliação;.

- a última, utilizando as seis variáveis testadas anteriormente, os três ponderadores e os três valores brutos da centralidade de grau, da distância geodésica até o grau 2 dos atores vinculados ao ator em avaliação e da recência dos atores vinculados ao ator em avaliação.

O procedimento “both” foi utilizado para selecionar as variáveis independentes que mais influenciam as variáveis dependentes, por implicar em um procedimento conservador, dado o duplo “filtro” das variáveis. O método dos mínimos quadrados (MMQ) foi utilizado para estimar os coeficientes de regressão em (1). Suponha que n > k observações são avaliadas, em que k é o número de variáveis, e y_i a i-ésima variável resposta observada e x_{ij} a i-ésima observação da j-ésima variável independente (i = 1, . . . , n, j = 1, . . . , k).

Assume-se que os erros ε_i são independentes e seguem distribuição normal com média zero e variância constante e desconhecida σ^2. Dado que os dados utilizados são sintéticos, não foram utilizados testes para verificação da hipótese de homocedasticidade.

4.4.1 Modelo de Regressão Linear Múltipla utilizando as Ponderações

Nesta etapa, buscou-se verificar a viabilidade da utilização das ponderações geradas na seção 4 para aprimorar a efetividade de um modelo de previsão. Sendo assim, para cada indivíduo “i”, foram consideradas a ponderação de centralidade de grau, PG_i, a ponderação da distância geodésica de até o grau 2, $PDG2_i$ e a ponderação pela recência, PR_i.

Buscou-se aqui fazer uma estimativa da pontuação via regressão linear múltipla, com

$$Y_i = I_0 + \sum_{i=1}^{N} a_i PG_i + \sum_{i=1}^{N} b_i PDG$$, onde I_0 é o intercepto.

As ponderações calculadas na seção anterior foram utilizadas para gerar o modelo de regressão linear múltipla, tendo sido definido o método “stepwise” com α para entrada = 0,15; α para remoção = 0,15, valores tipicamente utilizados.

Na rede social exemplo, a estrutura pré-regressão do modelo 1 de regressão está representada na tabela 2:

Tabela 2: Estrutura pré-regressão do modelo 1 na rede social exemplo

Y			
600,0			
350,0			
800,0			

Fonte: Elaboração do autor

4.4.2 Modelo de Regressão Linear Múltipla utilizando os Valores Brutos de Análise de Redes Sociais

Aqui buscou-se utilizar os valores “brutos” da análise de redes sociais como abordagem prospectiva. Sendo assim, para cada indivíduo “i”, foi considerada a sua própria centralidade de grau, $G_{i,}$, e, em relação aos demais atores que se vinculam ao ator em avaliação, a média da distância geodésica de até o grau 2, $DG2_i$ e a média da recência, $R_{i.}$

Também aqui foi aplicada uma regressão linear múltipla, com

$$Y_i = I_0 + \sum_{i=1}^{N} a_i G_i + \sum_{i=1}^{N} b_i DG2_i$$

onde I_0 é o intercepto.

Aqui também foi definido o método “stepwise” com α para entrada = 0,15; α para remoção = 0,15.

Na rede social exemplo, a estrutura pré-regressão do modelo 2 de regressão está representada na tabela 3:

Tabela 3: Estrutura pré-regressão do modelo 2 na rede social exemplo

Y			
600,0			
350,0			
800,0			

Fonte: Elaboração do autor

4.4.3 Modelo de Regressão Linear Múltipla utilizando todas as Variáveis Pesquisadas

Nesta etapa, deu-se prosseguimento à abordagem prospectiva. Sendo assim, para cada indivíduo “i”, foram consideradas as ponderações e os dados brutos da rede social utilizada.

Mais uma vez foi aplicada a regressão linear múltipla, com

$$Y_i = I_0 + \sum_{i=1}^{N} a_i PG_i + \sum_{i=1}^{N} b_i PDC$$

onde I_0 é o intercepto.

Assim, todas as variáveis utilizadas nas seções anteriores foram testadas para gerar o modelo de regressão linear múltipla, tendo sido também definido o método “stepwise” com α para entrada = 0,15; α para remoção = 0,15.

Na rede social exemplo, a estrutura pré-regressão do modelo 3 de regressão estaria representada na tabela 4:

Tabela 4: Estrutura pré-regressão do modelo 3 na rede social exemplo

Y						
600,0						
350,0						
800,0						

Fonte: Elaboração do autor

4.5 Mensuração da efetividade dos modelos de predição

Neste ponto do trabalho foi importante avaliar a efetividade dos modelos aplicados. Imagine-se uma formulação perfeita: neste caso não haveria diferença entre os valores efetivos registrados no birô de crédito e aquele obtido pela ponderação aplicada. Sendo assim, configura-se a aplicação de testes estatísti-

cos não paramétricos, já que não há alguma característica numérica a ser avaliada, e de maneira ideal, que este fosse baseado em postos (“ranking”), já que seria independente dos valores obtidos. No caso ideal, se os indivíduos fossem ordenados de forma crescente pela pontuação de score de birô real e depois pela pontuação inferida, o 1º indivíduo na lista do score real também seria o 1º na lista do score previsto, o que se repetiria para o 2º, assim como também ocorreria para o n-ésimo.

4.5.1 Coeficiente de Correlação Rhô de Spearman - ρ

O coeficiente de correlação de Spearman é um teste não paramétrico de postos. Esta estatística é a mais antiga e também a mais conhecida para variáveis mensuradas em nível ordinal, também denominado Coeficiente de Correlação por Postos de Spearman, e representado por ρ. Foi escolhido pelo fato de ser um teste em nível ordinal, o que o torna menos sensível a oscilações nos valores brutos de score e pontuações. Enfatize-se que as correlações ordinais não podem ser interpretadas da mesma maneira que a correlação de Pearson, que é uma medida do grau de relação linear entre duas variáveis quantitativas. Inicialmente, não mostram necessariamente tendência linear, mas podem ser consideradas como índices de monotonicidade, ou seja, para aumentos positivos da correlação, aumentos no valor de X correspondem a aumentos no valor de Y, e para coeficientes negativos ocorre o oposto. O quadrado do índice de correlação não pode ser interpretado como a proporção da variância comum às duas variáveis.

4.5.1.1 Estimador do Coeficiente de Correlação de Spearman

Este estimador surgiu como analogia ao estimador do Coeficiente de Correlação Linear de Pearson, conforme apresentado por Siegel (1975). Uma das formas de se apurar o valor da estimação é:

$$\rho = 1 - 6 \frac{\sum_{i-1}^{n} d_i^2}{n^3 - n}$$

Onde “n” é o número de pares (x_i, y_i);, sendo “x” o score de birô e “y” a previsão do modelo d_i =(posto de x_i dentre os valores de x) - (posto de y_i dentre os valores de y). Se os postos de x são exatamente iguais aos pontos de y, então todos os d_i serão zero e ρ será 1, onde o modelo seria “perfeito”.

O coeficiente ρ de Spearman varia entre -1 e 1. Quanto mais próximo estiver destes extremos, maior será a correlação entre as variáveis. O sinal negativo para esta correlação significa que as variáveis variam em sentido contrário, isto é, as categorias mais elevadas de uma variável estão associadas a categorias mais baixas da outra variável.

4.5.1.2 Avaliação da significância do Coeficiente de Correlação de Spearman

Para tal se faz necessária a realização de um teste de hipóteses, monocaudal, já que busca-se avaliar se os valores reais do score e as inferências confluem. Assim tem-se:

- Hipótese nula (H_0): não há associação entre o pseudo score de birô e a ponderação em avaliação.
- Hipótese alternativa (H_1): há associação entre o pseudo score de birô e a ponderação em avaliação.

O procedimento envolve a comparação do valor calculado da estatística ρ com o valor tabelado, que depende do número de pares (x,y) e do nível de significância adotado. Se o valor calculado for maior ou igual ao tabelado, rejeita-se H_0 (RAMSEY, 1989). Se o valor calculado for menor que valor tabelado, não se rejeita H_0. No limite, caso as duas distribuições implicassem em um mesmo ranking, não se poderia rejeitar a hipótese que elas tivessem uma mesma natureza estatística

4.5.2 O Teste χ^2

O teste χ^2 é utilizado para se averiguar a aderência de um conjunto de dados amostrais a uma dada distribuição de probabilidades ou de frequências. Assim, busca-se comprovar se existe diferença significativa entre o número observado de indivíduos, ou de respostas, em determinada categoria, e o respectivo valor esperado.

Karl Pearson (PLACKETT, 1983) propôs a seguinte fórmula para medir as possíveis discrepâncias entre proporções observadas e esperadas:

$$\chi^2 = \sum_{i=1}^{n} \left[\left(o_i - e_i \right)^2 / e_i \right]$$

Neste trabalho:

• "o_i" = frequência observada para cada classe, valor da pontuação de crédito calculado pelos três algoritmos;

• "e_i" = frequência esperada para aquela classe, valor original da pontuação de crédito.

Assim, quando as frequências observadas são relativamente próximas as esperadas, o valor de χ^2 é baixo. Mas, quando as divergências relativas são apreciáveis, $(o_i - e_i)/ e_i)$ se eleva e, consequentemente, χ^2 assume valores altos.

São testadas duas hipóteses:

• Hipótese nula (H_0): Não existe diferença relativa entre as frequências apuradas pelo modelo adotado, confrontadas com as esperadas na distribuição do score de birô de crédito, o que sugere eficiência do modelo;

• Hipótese alternativa (H_1): Há diferença relativa entre as frequências apuradas pelo modelo adotado, confrontadas com as esperadas na distribuição do pseudo score de birô de crédito.

O procedimento envolve a comparação do valor calculado da estatística χ^2 com o valor tabelado, que depende do número de graus de liberdade e do nível de significância adotado. Se o valor calculado for maior ou igual ao tabelado, rejeita-se

H_0. Se o valor calculado for menor ao valor tabelado, não se rejeita H_0.

4.6 Softwares utilizados

Foram empregados os softwares:

- Ucinet 6 para a Análise de Redes Sociais (https://sites.google.com/site/ucinetsoftware/downloads) da empresa Analytictech. A comunidade utilizada no trabalho faz parte dos conjuntos de dados fornecidos pelo provedor do software;
- MINITAB (www.minitab.com/), software para a aplicação de análises estatísticas;
- Planilhas Excel para a apuração dos valores ponderados de pontuação de crédito.

5. RESULTADOS

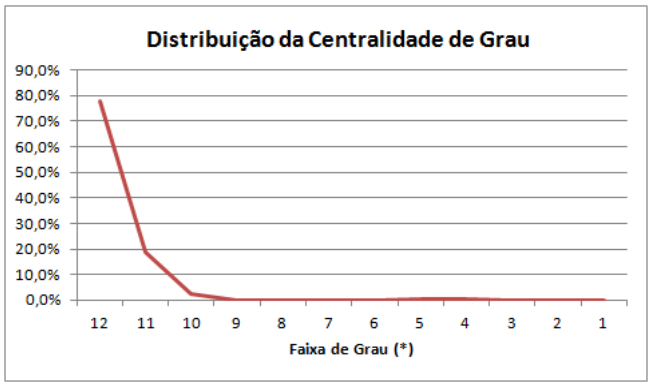
5.1 Análise da Rede Social utilizada

A seguir são apresentadas algumas características apuradas nesta rede de relacionamentos:

- Centralidade média de Grau (ou Grau médio): 48,790 - cada indivíduo se relaciona diretamente em média com 40 outros atores;
- Densidade: 5,1% - são estabelecidos 1 em cada 20 pares de vínculos possíveis entre os atores;
- Distância Geodésica média (ou distância média): 2,405 - em média são necessários 2 intermediários para colocar dois atores em contato;
- Desvio padrão da distância geodésica: 0,638 – cerca de 95% dos vínculos se dão entre 1 e 4 passos;
- Diâmetro: 5 – a maior distância geodésica possível entre dois indivíduos é de 5 passos;
- Tamanho da Matriz de Adjacência: 921.600 células

Esta Rede Social forma clusters, evidenciados pela concentração de atores em determinados setores, demonstrados na figura 7 e o pequeno diâmetro de cinco passos, ou graus, é representativa de um “Mundo Pequeno” (“Small World”), como comunidades fechadas de indivíduos. A figura 8 representa a distribuição da centralidade de grau, que parece seguir uma “Power Distribution”.

Figura 8: Distribuição do “Grau”* da Rede Social utilizada



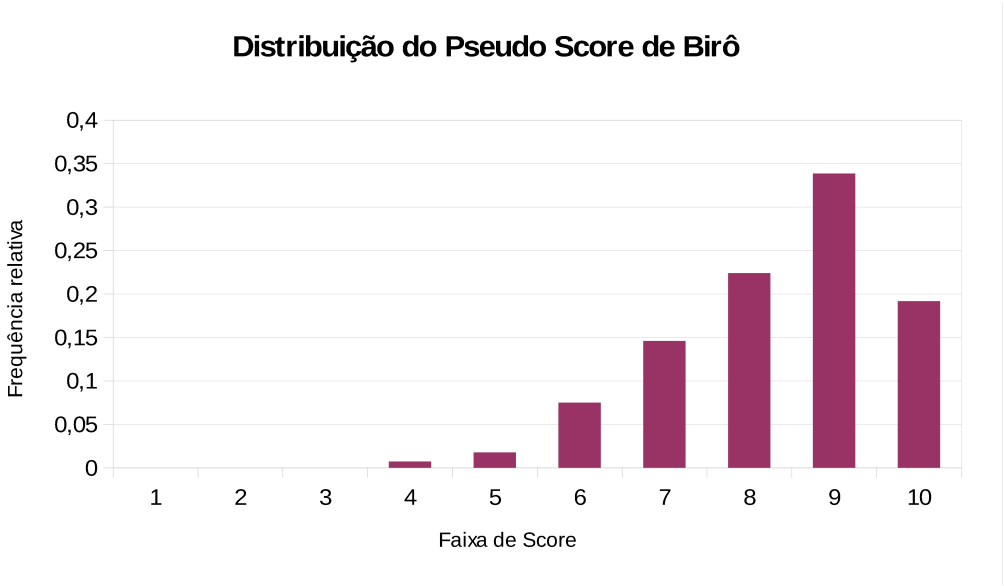
Fonte: Elaboração do autor

(*) Faixas de centralidade de grau, ou “Grau” – 12, de 250 a 300; 11, de 219 a 249; 10, de 213 a 218; 9, de 9 a 212, 8, com 8 indivíduos; 7, com 7 indivíduos e assim por diante

5.2. Geração de dados sintéticos de birôs de crédito via simulação estatística

A figura 9 ilustra a natureza da distribuição adotada visualmente se ajusta a uma distribuição Beta, com assimetria negativa.

Figura 9: A distribuição das pontuações do pseudo score de birô** de crédito gerada por simulação estatística.

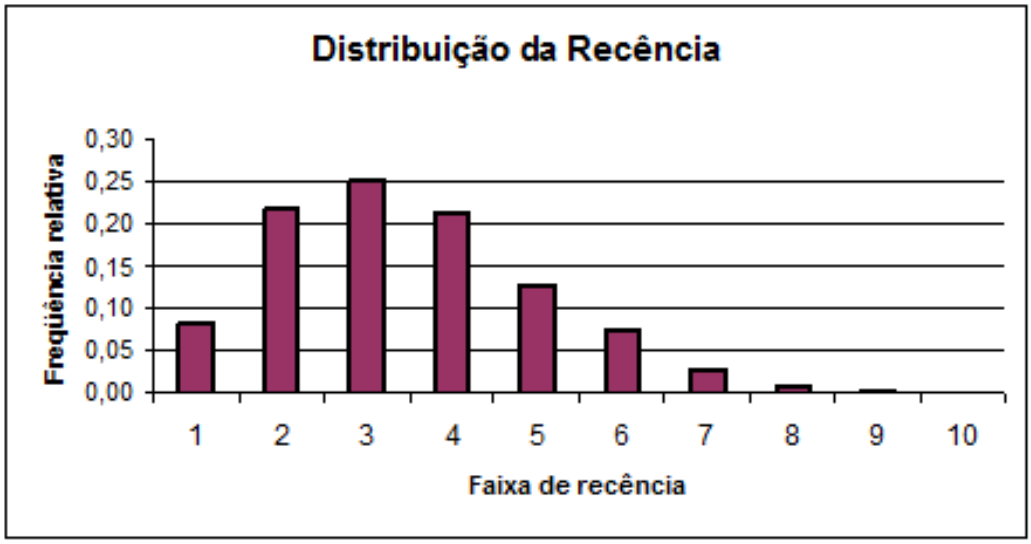


Fonte: Elaboração do autor

(**) Faixas do pseudo score - 1: de 0 a 99,9 pontos; 2, de 100 a 199,9: 3, de 200 a 299,9; 4, de 300 a 399,9; 5, de 400 a 499,9, 6, de 500 a 599,9; 7, de 600 a 699,9; 8, de 700 a 799,9; 9, de 800 a 899,9; 10: acima de 899,9

A figura 10, a seguir, revela que as faixas da recência** distribuíram-se de acordo com uma curva Weibull, como esperado.

Figura 10: A distribuição Recência gerada por simulação estatística



Fonte: Elaboração do autor

(**) Faixas de Recência - 1: até 1,9 anos; 2, de 2,0 a 3,9,; 3: de 4,0 a 5,9; 4, de 6,0 a 7,9; 5, de 8,0 a 9,9; 6, de 10,0 a 11,9; 7, de 12,0 a 13,9; 8, de 14,0 a 15,9; 9, de 16,0 a 17,9; 10: acima de 17,9 anos

5.3. Resultados dos modelos de ponderação

Nesta fase são apresentados os resultados dos três modelos de ponderação, pela centralidade de grau, pela "recência" e pela distância geodésica, conforme descritos na seção 4, avaliados conforme o critério de discriminação adotado. Neste contexto, ao contrário do que é praticado nos modelos de escoragem, o melhor modelo seria aquele em que as pontuações estimadas diferissem menos das pontuações de credit birô, que visualmente demonstrou se tratar da ponderação pela centralidade de grau. As tabelas 5 e 6 demonstram um sumário dos resultados dos testes estatísticos aplicados aos modelos de ponderação.

Enquanto para o teste "ρ" de Spearman diferenças estatisticamente significativas revelam possível associação entre as distribuições estatísticas, no teste "χ" quadrado, diferenças estatisticamente significativas apontam para fortes evidências de falta de associação entre as distribuições estatísticas confrontadas.

Tabela 5: Sumário dos resultados do teste estatístico “ρ” de Spearman

Modelo de Ponderação	“ρ” calculado	“ρ” crítico 1%	Diferença Significativa (s/n)	“ρ” crítico 5%	Diferença Significativa (s/n)
Grau	-0,0619	-0,0749	Não	-0,0533	**Sim**
Recência	0,0111	0,0749	Não	0,0533	Não
Distância Geodésica	-0,0135	-0,0749	Não	-0,0533	Não

Fonte: Elaboração do autor

Tabela 6: Sumário dos resultados do teste estatístico “χ Quadrado”

Modelo de Ponderação	“χ” calculado	“χ” crítico 1%	Diferença Significativa (s/n)	“χ” crítico 5%	Diferença Significativa (s/n)
Grau	9.312,29	21,67	**Sim**	16,92	**Sim**
Recência	954,96	21,67	**Sim**	16,92	**Sim**
Distância Geodésica	14.400,00	21,67	**Sim**	16,92	**Sim**

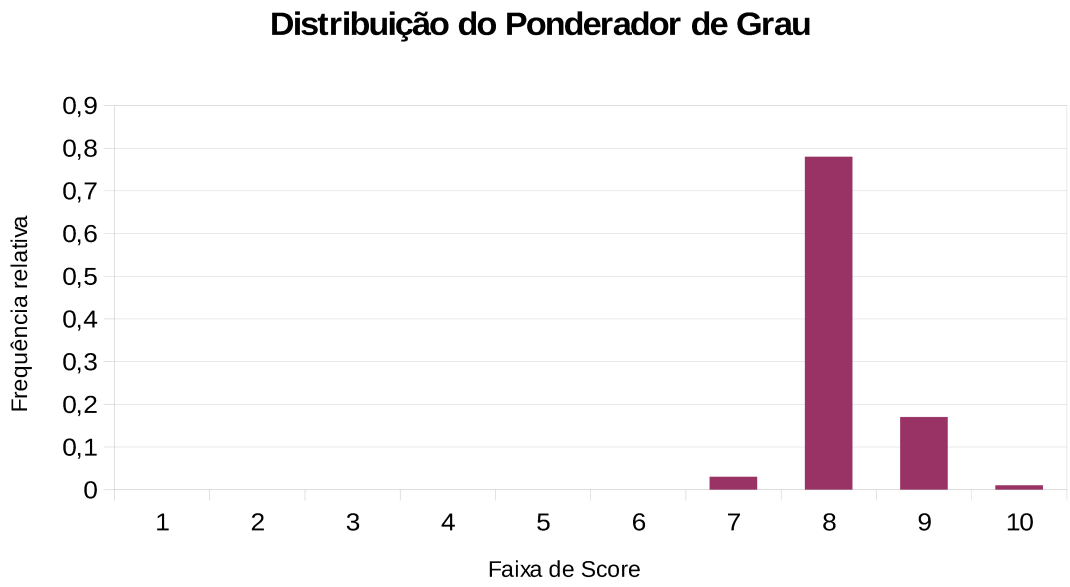
Fonte: Elaboração do autor

As seções a seguir revelam mais aspectos sobre o desempenho de discriminação dos modelos de ponderação.

5.3.1 Ponderação pela Centralidade de Grau

Os resultados desta fase estão apresentados na figura 11:

Figura 11: Distribuição das pontuações da ponderação por Grau *** (centralidade de grau)



Fonte: Elaboração do autor

(***) Faixas do score de ponderação - 1: de 0 a 99,9 pontos; 2, de 100 a 199,9: 3, de 200 a 299,9; 4, de 300 a 399,9; 5, de 400 a 499,9, 6, de 500 a 599,9; 7, de 600 a 699,9; 8, de 700 a 799,9; 9, de 800 a 899,9; 10: acima de 899,

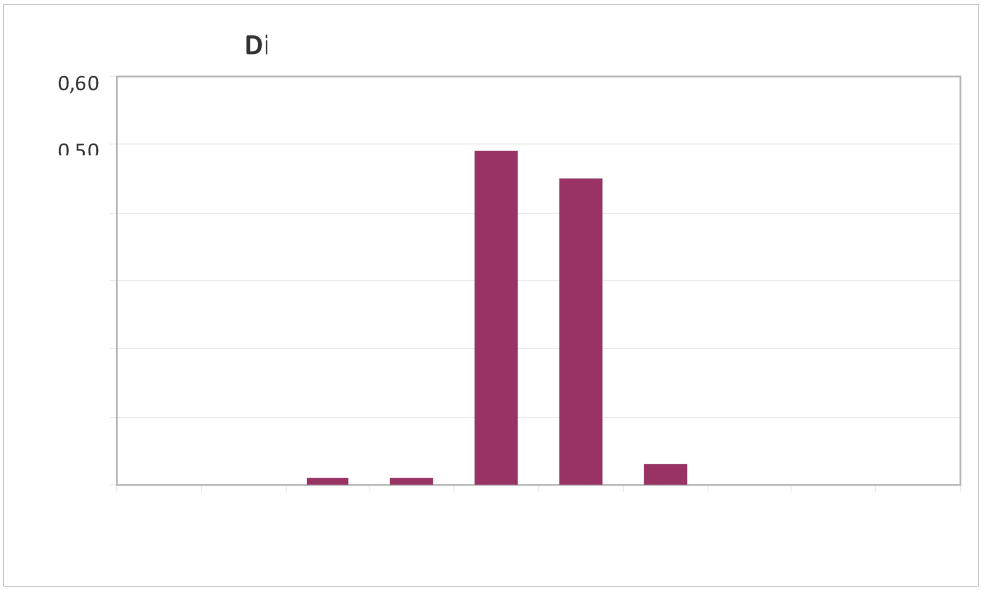
O valor calculado ρ-Spearman, da distribuição de frequências das pontuações obtidas pela ponderação de grau em relação ao pseudo score de birô, foi de -0,0619. Comparado aos valores críticos de ρ-Spearman a 1% ou a 5% de significância, -0,0749 e -0,0533, respectivamente, pode-se concluir que não se pode rejeitar a hipótese de que não existe associação entre as duas distribuições a 1% de significância, embora seja possível afirmar associação ao nível de 5%. Como os dados da distribuição do pseudo score foram gerados sinteticamente, o primeiro resultado pode ser considerado como flutuação estatística.

O valor da soma das distâncias “Qui-quadrado” em relação à distribuição do Pseudo score de Birô foi de 9.312,29 . Comparado ao valor crítico de Qui-quadrado a 5% ou a 1% de significância para 9 graus de liberdade, ou 10 classes, 16,92 e 21,67, respectivamente, pode-se concluir que há diferenças significativas entre as duas distribuições, como se esperava, pois os dados foram gerados sinteticamente.

5.3.2 Ponderação pela Recência

Os resultados desta etapa são apresentados na figura 12.

Figura 12: A distribuição das pontuações da ponderação pela Recência**** (tempo de relacionamento entre os atores)



Fonte: Elaboração do autor

(****) Faixas do score de ponderação - 1: de 0 a 99,9 pontos; 2, de 100 a 199,9: 3, de 200 a 299,9; 4, de 300 a 399,9; 5, de 400 a 499,9, 6, de 500 a 599,9; 7, de 600 a 699,9; 8, de 700 a 799,9; 9, de 800 a 899,9; 10: acima de 899,

O valor calculado ρ-Spearman da distribuição de frequências das pontuações obtidas pela ponderação de grau em relação ao pseudo score de birô foi de 0,0111. Comparado aos valores críticos de ρ-Spearman a 1% ou a 5% de significância, 0,0749 e 0,0533, respectivamente, pode-se concluir que não se pode rejeitar a hipótese que não existe associação entre as duas distribuições aos níveis de 1% e de 5% de significância. Como os dados da distribuição do pseudo score foram gerados sinteticamente, os dois resultados podem ser considerados como flutuação estatística.

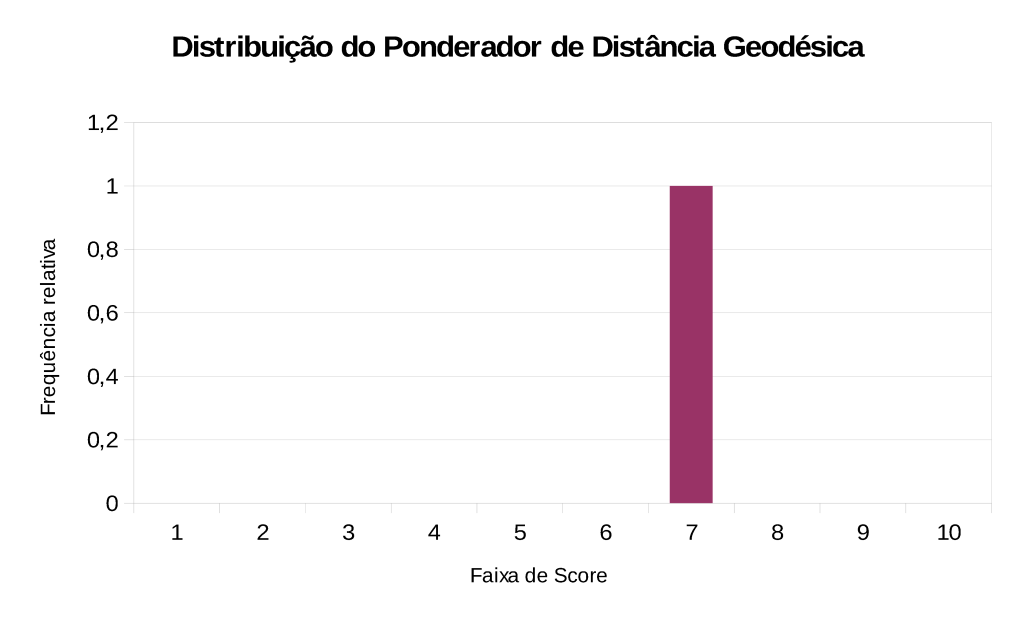
O valor da soma das distâncias “Qui-quadrado” em relação à distribuição do Pseudo score de Birô foi de 954,96. Comparado ao valor crítico de Qui-quadrado a 1% ou a 5% de significância para 9 graus de liberdade, ou 10

classes, 21,67 e 16,92 respectivamente, pode-se concluir que há diferenças significativas entre as duas distribuições, como se esperava, pois os dados foram gerados sinteticamente.

5.3.3 Ponderação pelas Distâncias Geodésicas

Os resultados deste modelo estão apresentados na figura 13.

Figura 13: A distribuição das pontuações da ponderação***** pelas distâncias geodésicas (menor caminho entre dois nós de um grafo)



Fonte: Elaboração do autor

(*****) Faixas do score de ponderação - 1: de 0 a 99,9 pontos; 2, de 100 a 199,9: 3, de 200 a 299,9; 4, de 300 a 399,9; 5, de 400 a 499,9, 6, de 500 a 599,9; 7, de 600 a 699,9; 8, de 700 a 799,9; 9, de 800 a 899,9; 10: acima de 899,

A concentração em uma só faixa pode ser explicada pelo Teorema Central do Limite, pois, em média, cada ator apresentou distâncias geodésicas 1 e 2 com 920 outros indivíduos. Tal efeito apresenta mesma natureza que o cálculo da média amostral, quando o desvio padrão desta estatística cai com o inverso do quadrado do número de elementos da amostra (MARTINS e DOMINGUES, 2011).

O valor calculado ρ-Spearman, da distribuição de frequências das pontuações obtidas pela ponderação de grau em relação ao pseudo score de

birô, foi de -0,0135. Comparado aos valores críticos de ρ-Spearman a 1% ou a 5% de significância, -0,0749 e -0,0533, respectivamente, pode-se concluir que não se pode rejeitar a hipótese de que não existe associação entre as duas distribuições aos níveis de 1% e de 5% de significância. Como os dados da distribuição do pseudo score foram gerados sinteticamente, os dois resultados podem ser considerados como flutuação estatística.

O valor da soma das distâncias "Qui-quadrado" em relação à distribuição do Pseudo score de Birô foi de 14.400,00. Comparado ao valor crítico de Qui-quadrado a 1% ou a 5% de significância para 9 graus de liberdade, ou 10 classes , 21,67 e 16,92 respectivamente, pode-se concluir que há diferenças significativas entre as duas distribuições, como se esperava, pois os dados foram gerados sinteticamente.

5.4 Modelos de Regressão Linear Múltipla

Os resultados dos três modelos desenvolvidos estão apresentados nas seções a seguir.

5.4.1 Modelo de Regressão Linear Múltipla utilizando as Ponderações

A equação de regressão obtida é apresentada a seguir:

PSB = 201,7 + 0,738 PG , onde,

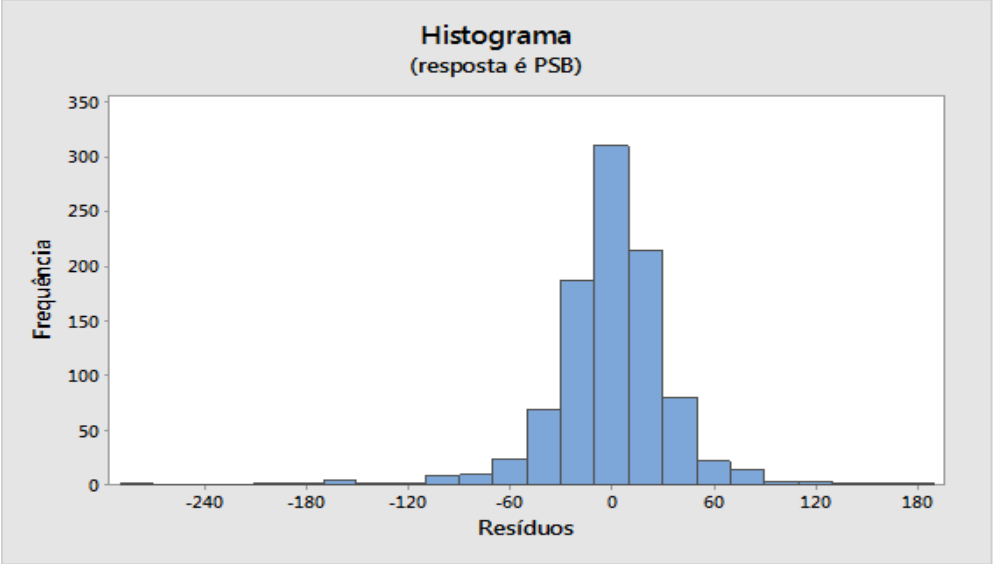
- PSB é a estimativa do pseudo score de birô e
- PG é a ponderação de grau.

Como já era esperado, com o uso dos dados sintéticos de pseudo score de birô de crédito, as estatísticas da análise de variância da regressão (R^2: 4.18%, R^2 Adjusted: 4,08%, F-statistic: 0.6948 on 494 and 5 DF, p-value: 0.7914), mostram que o modelo não pode ser considerado estatisticamente abrangente e estável, embora o p-valor da estatística "*f*" tenha sido nulo.

A análise dos resíduos da regressão mostra que a sua distribuição se concentrou próxima ao valor 0 e que estes apresentam distribuição que

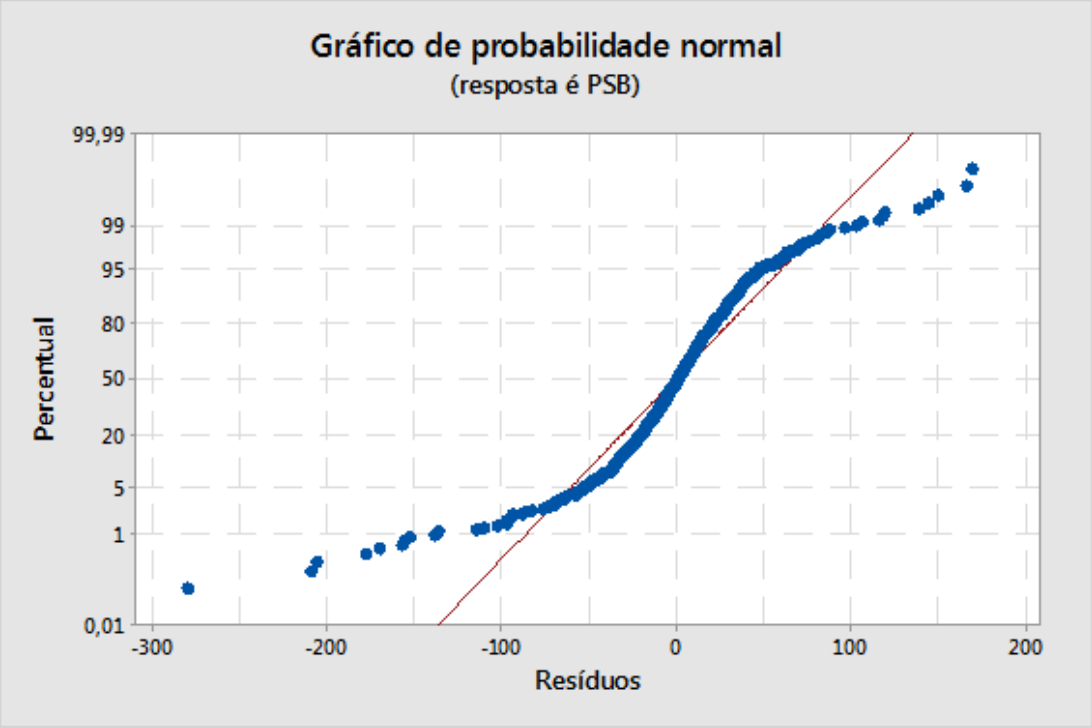
assemelha-se à Normal – espera-se uma reta com inclinação positiva, um dos requisitos básicos da regressão linear, conforme apontado respectivamente pelas figuras 14 e 15.

Figura 14: Histograma dos resíduos do modelo de regressão dos ponderadores



Fonte: Elaboração do autor

Figura 15: Distribuição dos resíduos do modelo de regressão dos ponderadores com escala Normal



Fonte: Elaboração do autor

5.4.2 Modelo de Regressão Linear Múltipla utilizando os Valores Brutos de Análise de Redes Sociais

Não houve variável que atendesse os requisitos do método stepwise e, assim, não houve regressão.

5.4.3 Modelo de Regressão Linear Múltipla utilizando todas as Variáveis Pesquisadas

A equação de regressão do 3º modelo é exatamente igual a do modelo 1, e é apresentada a seguir:

PSB = 201,7 + 0,738 PG , onde,

- PSB é a estimativa do pseudo score de birô e

- PG é a ponderação de grau.

As estatísticas da análise de variância da regressão também foram as mesmas do modelo de regressão com os ponderadores (R^2: 4.18%, R^2 Adjusted: 4,08%, F-statistic: 0.6948 on 494 and 5 DF, p-value: 0.7914), e mostram que o modelo não pode ser considerado estatisticamente abrangente e estável, embora o p-valor da estatística "*f*" tenha sido nulo.

Da mesma forma, a análise dos resíduos da regressão apresentou os mesmos resultados que anteriormente, mostrando que a sua distribuição se concentrou próxima ao valor zero, com distribuição semelhante à Curva Normal.

A tabela 7 sumariza os resultados apresentados nesta seção.

Tabela 7: Sumário dos resultados dos modelos de regressão linear múltipla

Modelo					
1					
2					

Fonte: Elaboração do autor

6. DISCUSSÃO E CONSIDERAÇÕES FINAIS

6.1. Conclusões

Pode-se considerar que o objetivo geral deste trabalho, que foi ampliar os conhecimentos sobre fontes e métodos alternativos de dados para suportar os sistemas e processos de decisão de crédito, com o intuito de aprimorar a eficácia dos modelos de apoio a decisão de crédito, foi atendido. Dados oriundos de redes sociais a priori dificultam a eventual manipulação de interessados para aumentar as chances de aprovação das solicitações de crédito, e os resultados apontam que tais dados aliados aos conjuntos de dados atualmente, deve propiciar maior efetividade dos modelos de apoio à decisão.

Ressalte-se que, em geral, os resultados dos testes de efetividade não revelaram aderência dos modelos às distribuições de probabilidade do pseudo score de birô de crédito, principalmente em decorrência da necessidade de se criarem dados sintéticos, que adicionalmente pertenciam à uma comunidade de indivíduos de fora do Brasil. Esta configuração se deu pela indisponibilidade para este trabalho, tanto de dados reais de birô positivo, quanto de dados de indivíduos brasileiros ou residentes no País de uma rede social livre de escala. Neste caso a expectativa é que os resultados devam apontar aderência estatisticamente aceitável das distribuições das estimativas com relação ao score de birô.

Entretanto, foram utilizadas neste trabalho técnicas que sinalizam potencial de utilização dos dados de redes sociais no aperfeiçoamento dos sistemas de decisão, um dos quais a Análise de Crédito. Mecanismos de pequena complexidade como ponderações utilizando as principais métricas, como centralidade de grau, distância geodésica e um novo conceito trazido por este trabalho, a “recência”, apontam, a princípio, potencial de utilização nos mecanismos de avaliação de crédito, possibilitando a medição objetiva dos mecanismos desenvolvidos e, dadas as semelhanças com outras áreas funcionais que se baseiam no comportamento dos indivíduos, trazem perspectivas de aplicação em Marketing, Vendas e a Gestão de Recursos Humanos.

Embora ainda não se tenham utilizados dados reais para a aplicação de mecanismos consagrados como a regressão linear múltipla, não se pode rejeitar a possibilidade da aplicação deste tipo de abordagem a redes sociais do tipo “mundos pequenos”. Pode-se elaborar a hipótese que quando um indivíduo novo for aceito pelos critérios de seleção da comunidade, seu perfil será próximo ao dos demais atores.

Adicionalmente, cabe ressaltar que houve apreciável dificuldade no acesso e manipulação de uma rede social de pequena dimensão, ponto que pode ser superado pelo uso de tecnologia de “big data”, que não estava acessível no contexto deste trabalho. Entretanto, foi demonstrada a viabilidade de acesso aos dados de redes sociais, que podem ser enriquecidos pelas informações usualmente tratadas pelos departamentos de “business inteligence” e de modelagem matemática, que, aliadas às ferramentas atualmente disponíveis para o tratamento de volumes astrônomicos de dados, devem ampliar sobremaneira as perspectivas de obtenção de resultados de previsão de comportamento sócio-econômico cada vez mais precisos e robustos.

Quanto aos objetivos específicos, foi atendido o propósito de apontar métodos de confronto do desempenho das estimativas em relação ao pseudo score de birô, mas os demais alvos não foram alcançados pelas mesmas razões que afetaram os resultados do objetivo geral.

Entretanto, o presente trabalho mostrou que já a partir de agora, as organizações que enfrentam o desafio de prever o comportamento de indivíduos e consumidores, podem se valer da ferramenta de ponderação e de médias móveis para incorporar o acervo informacional contido nas redes sociais aos seus mecanismos de decisão, operativos e de controle. Adicionalmente, cite-se a possibilidade de uso de dados de modelagem originadas por redes sociais ao já existente arsenal de opções para pautar decisões, como as matrizes de decisão (XU, 2009).

Estudo efetuado por pesquisadores com várias comunidades colombianas – de cultura latina e em ambiente similar ao nosso - expostos a situações de risco (ATTANASIO, BATTISTIN e MESNARD, 2012), apontou que há vantagens para os indivíduos que se agrupam colaborativamente, e que essas vantagens podem se tornar inacessíveis quando o grau de confiança

entre os indivíduos está ausente ou baixo. A “confiança” foi mensurada utlizando-se a análise de redes sociais. Pode-se especular, dado o presente estudo e aquele recém mencionado, que em situações de crise de inadimplência estrutural e disseminada, indivíduos com mais e mais consistentes laços sociais possam enfrentar mais a contento tais desafios e apresentarem comportamento mais favorável na recuperação de sua eventual inadimplência.

Bakshy et al (2012) demonstraram que apesar de laços fortes serem isoladamente mais influentes, são menos frequentes, e que os mais abundantes laços fracos acabam sendo responsáveis pela propagação de novas informações em redes sociais. Este fato sugere que estes laços mais fracos podem desempenhar um papel mais preponderante na divulgação de informações on-line do que atualmente se acredita. Trazendo para o escopo deste trabalho, o fluxo de informações pode ser um próxi do mecanismo de influência, sugerindo que quanto mais atores e vínculos forem observados, maior deverá ser o poder preditivo de modelos futuramente desenvolvidos.

Para a evolução deste trabalho, fica aqui o registro de etapas que apresentam potencial de superação dos obstáculos enfrentados até aqui, para a ampliação do escopo de pesquisa:

- firmar convênio com os birôs de crédito nacionais para a coleta de dados reais de scores e do maior conjunto possível de variáveis de natureza comportamental ou cadastral;

- estabelecer acordo com os provedores de redes sociais livre de escala para coleta e armazenagem de amostras de dados de indivíduos reais residentes no Brasil. Nesta etapa deverão estar superados os empecilhos jurídicos de utilização destes dados, possivelmente solucionados com termos de confidencialidade;

- contar com estrutura de hardware e de software com dimensionamento suficiente para processar volumes astronômicos de dados, conhecido como ferramental “big data”.

6.2. Limitações

As questões de indisponibilidade da informação do conceito de birô positivo, ainda em maturação e crescimento no País, do uso de dados de uma comunidade de indivíduos do exterior do País e da geração de dados sintéticos de “recência”, dada sua ausência na rede social utilizada, o poder de inferência das ferramentas foi decisiva e negativamente afetado;

O tratamento de dados de redes sociais exige estrutura robusta e sofisticada para gestão de “big data”, pois cada indivíduo gera quantidades extremamente grandes de dados, que devem ser condensados em informações relevantes e acessíveis. Novas pesquisas, acadêmicas ou comerciais, devem colher resultados ainda mais auspiciosos, caso contem com tal infra-estrutura.

Outro o ponto a ser considerado é o fato de a rede social adotada apresentar configuração de “mundo pequeno”, que se adéqua melhor aos modelos utilizados em abordagens ativas de empresas, quando estas ofertam crédito aos indivíduos e empresas, que usualmente utilizam nichos de mercado, como assinantes de determinados serviços ou membros de determinadas associações ou agremiações com perfil mais uniforme e maior densidade de relacionamento entre os atores membros. A despeito de este ser um importante nicho e de resultados econômico-financeiros bastante expressivos em operações financeiras, a abordagem “aberta” de crédito, em que os entes procuram as empresas para se candidatarem a crédito, não está coberta neste escopo. Esta configuração é mais adequada às redes “livres de escala”, em que as distâncias geodésicas são em média maiores[1].

---

1 Os conjuntos de dados avaliados possuíam tamanho mais reduzido que o PV960, o que limitaria em muito a representatividade do ferramental estatístico utilizado e dos resultados e conclusões obtidas

6.3. Recomendações para trabalhos futuros

Um passo natural para complementar o escopo deste trabalho é o de utilizar uma rede social “Livre de Escala” como fonte de dados (WATTS, STROGATZ, 1998), que poderiam ser enriquecidos com dados de credi birô positivo. Possivelmente não haveria a ocorrência de falta de ajuste dos modelos estatísticos ao grupo de controle.

Outra possibilidade é a utilização de métricas de redes sociais não empregadas no presente trabalho, como “centralidade de intermediação”, “centralidade de proximidade”, “centralidade de vetor próprio” e “força das ligações”, por exemplo, (TOMAÉL, MARTELETO, 2005).

Adicionalmente, há potencial de melhoria de resultados pela aplicação de ferramentas de contexto, como Redes Neurais e Cadeias de Markov, como alternativas para a modelagem matemática.

REFERÊNCIAS

ALTMAN, E.; HALDEMAN, R.; NARAYANAN, P. ZETA TM analysis: a new model to identify bankruptcy risk of corporations. Journal of Banking and Finance, Amsterdam, v. 1, n. 1, p. 29-54, 1977.

ATTANASIO, O.; BATTISTIN, E.; MESNARD, A. Food and cash transfers: Evidence from Colombia*. The Economic Journal, Chichester, v. 122, n. 559, p. 92-124, 2012.

BAKSHY, E. et al. The role of social networks in information diffusion. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 21., 2012, Lyon. Proceedings… Lyon: ACM, 2012. p. 519-528.

BARABÁSI, A. Linked: how everything is connected with everything else and what it means for business, science and everyday life. Cambridge: Plume, 2002. 285 p.

BARABÁSI, A-L.; ALBERT, R. Emergence of scaling in random networks. Science, New York, v. 286, n. 5439, p. 509-512, 1999.

BARABÁSI, A-L.; BONABEAU, E. Scale-free networks. Scientific American, New York, v. 5, n. 288, p. 60-9, 2003.

BENDER, R.; AUGUSTIN, T.; BLETTNER, M. Generating survival times to simulate Cox proportional hazards models. Statistics in medicine, Chichester, v. 24, n. 11, p. 1713-1723, 2005.

CHANDLER, G.; EWERT, D. Discrimination on the basis of sex under the Equal credit opportunity act. West Lafayette: Krannert Graduate School of Management, Purdue University, 1976.

DANYLLO, W. et al. Identifying relevant users and groups in the context of credit analysis based on data from Twitter. In: INTERNATIONAL

CONFERENCE ON CLOUD AND GREEN COMPUTING, 3., 2013, Karlsruhe. Proceedings… Karlsruhe: IEEE, 2013. p. 587-592.

DIXON, N. Common knowledge: how companies thrive by sharing what they know. Boston: Harvard Business Press, 2000.

ERDÖS, P.; RÉNYI, A. On the evolution of random graphs. Publication of the Mathematical Institute of the Hungarian Academy Science, Budapest, v. 5, p. 17-61, 1960.

FEDERAÇÂO BRASILEIRA DE BANCOS (FEBRABAN). Relatório Anual 2013. Disponível em: <http://www.febraban.org.br/downloads/RelatorioAnual/Relatorio_Anual_FEBRABAN_2013.pdf>. Acesso em: 20 jul. 2014.

FEI, H. et al. Content based social behavior prediction: a multi-task learning approach. In: CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 20., 2011, Glasgow. Proceendings… Glasgow: ACM, 2011. p. 995-1000.

FREEMAN, L. C. Centrality in social networks conceptual clarification. Social networks, Lausanne, v. 1, n. 3, p. 215-239, 1979.

HAND, D.; HENLEY, W. Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), London, v. 160, n. 3, p. 523-541, 1997.

HANNEMAN, R.; RIDDLE, M. Social network data: In: _____.Introduction to social network methods. Riverside: University of California, 2005. p. 4-22.

KRAMER, A.; GUILLORY, J.; HANCOCK, J. Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, Hanover, v. 111, n. 24, p. 8788-8790, 2014.

LEVIN, D.; CROSS, R. The strength of weak ties you can trust: the mediating role of trust in effective knowledge transfer. Management science, Providence, v. 50, n. 11, p. 1477-1490, nov. 2004.

MARTINS, G.; DOMINGUES, O. Estatística geral e aplicada. 4. ed. São Paulo: Atlas, 2011.

McPHERSON, M.; SMITH-LOVIN, L.; COOK, J. Birds of a feather: homophily in social networks. Annual Review of Sociology, Palo Alto, v. 27, p. 415-444, 2001.

NIELSEN IBOPE Número de pessoas com acesso a internet no Brasil supera 120 milhões. 2014. Disponível em: <http://www.nielsen.com/br/pt/press-room/2014/Numero-de-pessoas-com-acesso-a-internet-no-Brasil-supera-120-milhoes.html/>. Acesso em: 21 jul. 2014.

OPSAHL, T.; AGNEESSENS, F.; SKVORETZ, J. Node centrality in weighted networks: generalizing degree and shortest paths. Social Networks, Lausanne, v. 32, n. 3, p. 245-251, 2010.

PLACKETT, L. Karl Pearson and the chi-squared test. International Statistical Review/Revue Internationale de Statistique, The Hague, p. 59-72, 1983.

PRADO, R.; BASTOS, N.; DUARTE JR, A. Gerenciamento de riscos de crédito em bancos de varejo no Brasil. Tecnologia de Crédito, São Paulo, n. 65, p. 30-58, jul./ago. 2003.

RAMSEY, H. Critical values for Spearman's rank order correlation. Journal of Educational and Behavioral Statistics, Washington, v. 14, n. 3, p. 245-253, 1989.

SABATER, J.; SIERRA, C. Reputation and social network analysis in multi-agent systems. In: INTERNATIONAL JOINT CONFERENCE ON AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS: part 1, 1., 2002, Bologna. Proceedings… New York: ACM, 2002. p. 475-482.

SIEGEL, S. Estatística não-paramétrica para as ciências do comportamento. São Paulo: McGraw-Hill, 1975.

SCOTT, J. The development of social network analysis. In: _____. Social network analysis: a handbook. 2nd. ed. London: SAGE, 1999. p. 7-33.

SONG, M. et al. Construction and adoption of Net-enabled Credit Analysis Supporting System (NCASS) using social information networks. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL ASPECTS OF SOCIAL NETWORKS, 10., 2010, Taiwan. Proceedings... Taiwan: IEEE, 2010. p. 460-463.

TABACHNICK, G. et al. Using multivariate statistics. 4th ed. Boston: Allyn and Bacon, 2001.

THOMAS, L.; EDELMAN, D.; CROOK, J. Measuring scorecard performance. In: _____. Credit scoring and its applications. Philadelphia: SIAM, 2002. p. 107-117.

TOMAÉL, M. I.; MARTELETO, R. M. Redes sociais: posições dos atores no fluxo da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Florianópolis. Anais... Florianópolis, SC: UFSC. Disponível em: <http://repositorios.questoesemrede.uff.br/repositorios/bitstream/handle/123456789/327/gt3_tomael_marteleto.pdf?sequence=1>. Acesso em: jun. 2013.

WASSERMAN, S.; FAUST, K. Social network analysis: methods and applications. Cambridge: Cambridge University Press, 1994. 867 p.

WATTS, D.; STROGATZ, S. Collective dynamics of ‘small-world’ networks. Nature, London, v. 393, n. 6684, p. 440-442, 1998.

WYDICK, B.; KARP, H.; HILLIKER, K. Social networks, neighborhood effects, and credit access: evidence from rural Guatemala. World Development, Oxford,

v. 39, n. 6, p. 974-982, 2011.

XU, Z. An automatic approach to reaching consensus in multiple attribute group decision making. Computers & Industrial Engineering, Amsterdam, v. 56, n. 4, p. 1369-1374, 2009.

YU, Z.; YAN, H.; CHENG, T. Benefits of information sharing with supply chain partnerships. Industrial management & Data systems, [Bingley], v. 101, n. 3, p. 114-121, 2001.

ANEXO – SAÍDAS DO SOFTWARE MINITAB DOS TRÊS MODELOS DE REGRESSÃO LINEAR MÚLTIPLA

a) Modelo de Regressão Linear Múltipla utilizando as Ponderações

Resultados de: Arquivo regressão Ponderadores NEW.csv

Análise de Regressão: PSB versus DG2; Recência; Grau

Seleção Stepwise de Termos

α para entrada = 0,15; α para remoção = 0,15

Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Regressão	1	56019	56019	41,67	0,000
Grau	1	56019	56019	41,67	0,000
Erro	956	1285191	1344		
Total	957	1341210			

Sumário do Modelo

S	R2	R2(aj)	R2(pred)
36,6653	4,18%	4,08%	3,56%

Coeficientes

Termo	Coef	EP de Coef	Valor T	Valor-P	VIF
Constante	201,7	88,5	2,28	0,023	
Grau	0,738	0,114	6,46	0,000	1,00

Equação de Regressão

PSB = 201,7 + 0,738 Grau

Ajustados e Diagnósticos para Observações Atípicas

Obs.	PSB	Ajuste	Resid	Resid Pad		
20	828,00	794,39	33,61	0,92		X
22	685,00	781,43	-96,43	-2,63	R	
23	871,00	764,44	106,56	2,91	R	
34	829,00	791,36	37,64	1,03		X
35	847,00	773,26	73,74	2,01	R	
39	697,00	779,66	-82,66	-2,26	R	
47	659,00	755,46	-96,46	-2,64	R	X
60	776,00	748,17	27,83	0,76		X
76	863,00	777,58	85,42	2,33	R	
82	718,00	744,48	-26,48	-0,73		X
98	679,00	761,22	-82,22	-2,25	R	
121	610,00	787,06	-177,06	-4,84	R	
146	517,00	796,61	-279,61	-7,67	R	X
151	815,00	793,91	21,09	0,58		X
162	887,00	770,90	116,10	3,17	R	
200	692,00	768,40	-76,40	-2,09	R	
208	691,00	785,40	-94,40	-2,58	R	
216	649,00	742,06	-93,06	-2,56	R	X
265	901,00	781,66	119,34	3,26	R	

295	731,00	755,22	-24,22	-0,66		X
310	688,00	775,90	-87,90	-2,40	R	
316	877,00	780,37	96,63	2,64	R	
319	654,00	768,37	-114,37	-3,12	R	
350	889,00	769,87	119,13	3,25	R	
354	906,00	766,39	139,61	3,81	R	
372	847,00	743,20	103,80	2,86	R	X
376	796,00	809,81	-13,81	-0,38		X
381	755,00	754,22	0,78	0,02		X
404	870,00	782,41	87,59	2,39	R	
427	635,00	772,88	-137,88	-3,76	R	
457	777,00	793,40	-16,40	-0,45		X
470	717,00	750,92	-33,92	-0,93		X
496	930,00	784,79	145,21	3,97	R	
501	700,00	680,34	19,66	0,58		X
544	582,00	787,99	-205,99	-5,63	R	
556	764,00	742,16	21,84	0,60		X
557	826,00	801,78	24,22	0,67		X
559	630,00	765,50	-135,50	-3,70	R	
565	697,00	770,90	-73,90	-2,02	R	
568	850,00	761,98	88,02	2,40	R	
570	908,00	757,30	150,70	4,12	R	
572	839,00	762,21	76,79	2,10	R	
585	724,00	754,13	-30,13	-0,82		X
616	629,00	784,14	-155,14	-4,24	R	
622	828,00	794,38	33,62	0,92		X
628	849,00	774,22	74,78	2,04	R	
629	841,00	793,58	47,42	1,30		X
633	952,00	781,89	170,11	4,65	R	
639	550,00	759,11	-209,11	-5,72	R	
716	620,00	772,37	-152,37	-4,16	R	
735	831,00	794,22	36,78	1,01		X
738	670,00	779,33	-109,33	-2,98	R	
747	680,00	777,01	-97,01	-2,65	R	
753	800,00	812,46	-12,46	-0,34		X
774	849,00	766,87	82,13	2,24	R	
820	822,00	794,80	27,20	0,75		X
821	610,00	779,89	-169,89	-4,64	R	
826	857,00	774,50	82,50	2,25	R	
846	675,00	777,57	-102,57	-2,80	R	
850	775,00	793,59	-18,59	-0,51		X
851	776,00	748,17	27,83	0,76		X
857	931,00	764,82	166,18	4,54	R	
861	680,00	767,30	-87,30	-2,38	R	
862	670,00	763,43	-93,43	-2,55	R	
871	617,00	773,70	-156,70	-4,28	R	
919	832,00	794,18	37,82	1,04		X
923	813,00	805,51	7,49	0,21		X
952	853,00	772,06	80,94	2,21	R	

R Resíduo grande
X Atípicos X

b) Modelo de Regressão Linear Múltipla utilizando os Valores Brutos de Análise de Redes Sociais

Resultados de: NEW Brutos.csv

Análise de Regressão: PSB versus Grau; Média Rec; Média DG2

```
* NOTA * Não há termos no modelo.


Seleção Stepwise de Termos

α para entrada = 0,15; α para remoção = 0,15
Nenhum termo pode ser inserido no modelo.
```

c) Modelo de Regressão Linear Múltipla utilizando todas as Variáveis Pesquisadas

Resultados de: Arquivo regressão TODOS NEW.csv

Análise de Regressão: PSB versus DG2; Recência; Grau; Grau_1; Média Rec; Média DG2

```
Método

Linhas não usadas  2


Seleção Stepwise de Termos

α para entrada = 0,15; α para remoção = 0,15


Análise de Variância

Fonte       GL  SQ (Aj.)  QM (Aj.)  Valor F  Valor-P
Regressão    1     56019     56019    41,67    0,000
  Grau       1     56019     56019    41,67    0,000
Erro       956   1285191      1344
Total      957   1341210


Sumário do Modelo

      S     R2  R2(aj)  R2(pred)
36,6653  4,18%   4,08%     3,56%


Coeficientes

                 EP de
Termo       Coef   Coef  Valor T  Valor-P   VIF
Constante  201,7   88,5     2,28    0,023
Grau       0,738  0,114     6,46    0,000  1,00
```

Equação de Regressão

PSB = 201,7 + 0,738 Grau

Ajustados e Diagnósticos para Observações Atípicas

Obs.	PSB	Ajuste	Resid	Resid Pad		
20	828,00	794,39	33,61	0,92		X
22	685,00	781,43	-96,43	-2,63	R	
23	871,00	764,44	106,56	2,91	R	
34	829,00	791,36	37,64	1,03		X
35	847,00	773,26	73,74	2,01	R	
39	697,00	779,66	-82,66	-2,26	R	
47	659,00	755,46	-96,46	-2,64	R	X
60	776,00	748,17	27,83	0,76		X
76	863,00	777,58	85,42	2,33	R	
82	718,00	744,48	-26,48	-0,73		X
98	679,00	761,22	-82,22	-2,25	R	
121	610,00	787,06	-177,06	-4,84	R	
146	517,00	796,61	-279,61	-7,67	R	X
151	815,00	793,91	21,09	0,58		X
162	887,00	770,90	116,10	3,17	R	
200	692,00	768,40	-76,40	-2,09	R	
208	691,00	785,40	-94,40	-2,58	R	
216	649,00	742,06	-93,06	-2,56	R	X
265	901,00	781,66	119,34	3,26	R	
295	731,00	755,22	-24,22	-0,66		X
310	688,00	775,90	-87,90	-2,40	R	
316	877,00	780,37	96,63	2,64	R	
319	654,00	768,37	-114,37	-3,12	R	
350	889,00	769,87	119,13	3,25	R	
354	906,00	766,39	139,61	3,81	R	
372	847,00	743,20	103,80	2,86	R	X
376	796,00	809,81	-13,81	-0,38		X
381	755,00	754,22	0,78	0,02		X
404	870,00	782,41	87,59	2,39	R	
427	635,00	772,88	-137,88	-3,76	R	
457	777,00	793,40	-16,40	-0,45		X
470	717,00	750,92	-33,92	-0,93		X
496	930,00	784,79	145,21	3,97	R	
501	700,00	680,34	19,66	0,58		X
544	582,00	787,99	-205,99	-5,63	R	
556	764,00	742,16	21,84	0,60		X
557	826,00	801,78	24,22	0,67		X
559	630,00	765,50	-135,50	-3,70	R	
565	697,00	770,90	-73,90	-2,02	R	
568	850,00	761,98	88,02	2,40	R	
570	908,00	757,30	150,70	4,12	R	
572	839,00	762,21	76,79	2,10	R	
585	724,00	754,13	-30,13	-0,82		X
616	629,00	784,14	-155,14	-4,24	R	
622	828,00	794,38	33,62	0,92		X
628	849,00	774,22	74,78	2,04	R	
629	841,00	793,58	47,42	1,30		X
633	952,00	781,89	170,11	4,65	R	
639	550,00	759,11	-209,11	-5,72	R	
716	620,00	772,37	-152,37	-4,16	R	
735	831,00	794,22	36,78	1,01		X
738	670,00	779,33	-109,33	-2,98	R	
747	680,00	777,01	-97,01	-2,65	R	
753	800,00	812,46	-12,46	-0,34		X
774	849,00	766,87	82,13	2,24	R	
820	822,00	794,80	27,20	0,75		X
821	610,00	779,89	-169,89	-4,64	R	
826	857,00	774,50	82,50	2,25	R	
846	675,00	777,57	-102,57	-2,80	R	
850	775,00	793,59	-18,59	-0,51		X

851	776,00	748,17	27,83	0,76		X
857	931,00	764,82	166,18	4,54	R	
861	680,00	767,30	-87,30	-2,38	R	
862	670,00	763,43	-93,43	-2,55	R	
871	617,00	773,70	-156,70	-4,28	R	
919	832,00	794,18	37,82	1,04		X
923	813,00	805,51	7,49	0,21		X
952	853,00	772,06	80,94	2,21	R	

R Resíduo grande
X Atípicos X