UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS SOCIAIS APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO
CURSO DE MESTRADO EM ADMINISTRAÇÃO

RICARDO LUVIZOTTO DÓRIA

A DIFUSÃO DE INFORMAÇÃO EM REDE SOCIAL ONLINE: INFLUÊNCIA PESSOAL E PROPAGAÇÃO.

CURITIBA
2013

RICARDO LUVIZOTTO DÓRIA

A DIFUSÃO DE INFORMAÇÃO EM REDE SOCIAL ONLINE: INFLUÊNCIA PESSOAL E PROPAGAÇÃO.

Dissertação apresentada ao Programa de Pós-Graduação em Administração –PPGADM da Universidade Federal do Paraná – UFPR, como requisito parcial à obtenção do título de Mestre em Administração.

Orientador: Prof. Dr. Zaki Akel Sobrinho.
Co-orientação: Prof. Dr. Paulo Prado.

CURITIBA

2013

UNIVERSIDADE FEDERAL DO PARANÁ. SISTEMA DE BIBLIOTECAS.
CATALOGAÇÃO NA FONTE

Dória, Ricardo Luvizotto
A difusão de informação em rede social online: influência pessoal e propagação / Ricardo Luvizotto Dória. - 2013.
106 f.

Orientador: Zaki Akel Sobrinho.
Dissertação (Mestrado) – Universidade Federal do Paraná. Programa de Pós-Graduação em Administração, do Setor de Ciências Sociais Aplicadas.
Defesa: Curitiba, 2013.

1. Redes sociais on-line. 2. Disseminação da informação. 3.Twitter. 4. Internet. I. Akel Sobrinho, Zaki, 1957-. II. Universidade Federal do Paraná. Setor de Ciências Sociais Aplicadas. Programa de Pós-Graduação em Administração. III. Titulo.

CDD 658.4038

TERMO DE APROVAÇÃO

Ricardo Luvizotto Dória

"A DIFUSÃO DE INFORMAÇÃO EM REDE SOCIAL ONLINE: INFLUÊNCIA PESSOAL E PROPAGAÇÃO"

DISSERTAÇÃO APROVADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE NO PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO DA UNIVERSIDADE FEDERAL DO PARANÁ, PELA SEGUINTE BANCA EXAMINADORA:

Prof. Dr. Zaki Akel Sobrinho
(Orientador/UFPR)

Prof.ª Dr.ª Eliane Cristine Francisco-Maffezzolli
(Examinadora/PUC-PR)

Prof. Dr. Paulo Henrique Muller Prado
(Examinador/UFPR)

26 de agosto de 2013

DEDICATÓRIA

Dedico este trabalho ao meu avô, **Pedro Ricardo Dória**, pensador e cidadão do mundo.

AGRADECIMENTO

Ao **Alisson Prestes** (@javalisson), programador-hacker-ninja, pelo suporte tecnológico sem o qual esse trabalho não existiria.

Ao meus pai, **Ricardo José Dória**, pelo suporte e madrugadas de orientação e à minha mãe, Márcia Regina Luvizotto Dória por não ter me deixado desistir.

Ao professor **Dr. Paulo Henrique Mueller Prado** por ter acreditado mais em mim do que eu mesmo, oferecendo todo o suporte ao seu alcance para que eu pudesse concluir o trabalho.

À minha noiva **Flávia**, pela compreensão quando eu dividi o meu tempo entre o projeto "dissertação" e o projeto "nossa vida".

Ao Dream Team Aldeia: **Andreza**, **Letícia**, **Ivan**, **Kássia** e **Karen**, por deixar a Aldeia melhor a cada dia que eu passava fora.

Ao Dream Team Escola: **Guilherme**, **Fran**, **Joice**, **Yala** e **Thais**, pela compreensão e paciência na minha ausência.

RESUMO

O principal objetivo desta dissertação, apoiada na literatura de Difusão de Informações e das teorias de Análise de Estrutura de redes Sociais, foi de explorar a relação entre a difusão de uma informação na rede social *Twitter* e a estrutura da rede dos emissores da mensagem.

A partir de um processo de coleta de dados automatizado, foram analisados 846.441 membros da rede, totalizando 2.790 casos de difusão, divididos em 3 tópicos de discussão (*hashtags*). O caráter metodológico foi quantitativo não-probabilístico. Os resultados das analises de regressão aplicadas confirmaram 5 das 7 hipóteses de pesquisa para a amostra do trabalho.

Foi confirmado impacto de elementos da estrutura da rede na difusão de informações no Twitter, mas houve grande variação do impacto entre as hashtags analisadas. Estes resultados apontam a necessidade de novos estudos no campo.

Palavras-chave: Difusão de Informações, Estrutura de Redes Sociais, Computação de Big Data.

ABSTRACT

The main goal of this work, supported by the literature of Information Diffusion and Social Network Structure Analysis, was to explore the relationship between the diffusion of information on the Twitter social network and the structure of the network itself.

Through an automatic data collection process, 846.411 members of the network were analyzed. These members took part of 2.790 cases of diffusion, divided in 3 discussion topics (hash tags). The methodological character of this study was quantitative non-probabilistic. The results of the applied regression analysis confirmed 5 of the 7 research hypothesis for the sample.

The impact of the elements of the social network structure on the information diffusion on Twitter was confirmed, but a big variation in this impact among the analyzed hash tags was found. These results point the need for new studies in the field.

Key-words: Information Diffusion, Social Network Structure Analysis, Big Data Computing

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A difusão de novidades “boca a boca” é reconhecida como fator de influência de consumo e propagação de informações desde muito antes do surgimento da internet (Ryan and Gross, 1943). Estudos examinando redes sociais e seu papel na difusão de informação já foram realizados na antropologia, economia, geografia, sociologia e marketing (Hagerstrand, 1967; Robertson, 1971; Brown, 1981; Rogers, 2003) e, de alguma maneira, já foram adaptados de áreas como epidemiologia (Bailey, 1975).

Vista a importância da compreensão da difusão de informações para o lançamento de novos produtos, a literatura produzida pela área do marketing para esse campo é vasta e data de, ao menos, 1960. (Fourt and Woodlock, 1960) (Chandrasekaran e Tellis, 2007). Rogers (1976) apresentou a primeira moldura teórica do processo de adoção de inovações, apresentando o aspecto social como extremamente importante para qualquer tipo de processo de difusão em grupos sociais e Bass (1969) abriu o campo de estudos empíricos do processo de difusão, apresentando modelos matemáticos de mensuração e a dicotomia de papeis do inovador e imitador.

Nos últimos anos, percebeu-se, proveniente da massificação da utilização de computadores, telefones celulares, tablets e outros dispositivos eletrônicos na comunicação, o surgimento de novas tecnologias de interação social online, o que possibilitou o aumento do volume de troca de informações por meio de plataformas de *blogging* (como o Blogger.com), *micro-blogging* (como o *Twitter*.com), redes sociais (como o Facebook.com e *Twitter*.com) e mensagens instantâneas (como o Msn.com). À academia, estes avanços têm oferecido grandes subsídios para analisar a difusão em rede a partir de um volume de informação nunca antes visto. É possível, com alto nível de qualidade, mapear redes em grande escala e resgatar históricos de interações sociais ao longo do tempo.

Aos profissionais de marketing, as novas tecnologias têm se apresentado como uma oportunidade para formular estratégias e ações a fim de oferecer produtos e marcas nos canais onde estão os consumidores (Godin, 1999). Katona, Zubcsek e Sarvary (2010), no entanto, chamam atenção a uma mudança na perspectiva na inserção de anúncios publicitários: a utilização dos formatos

invasivos tipo “banner” têm provocado resultados desapontadores, enquanto aumentar o efeito do “boca a boca” tem sido uma maneira de utilizar as redes sociais para o marketing de maneira eficiente.

Os efeitos de viralização, apesar de ainda pouco controláveis e previsíveis, mostraram-se uma das maneiras mais eficientes de propagar informação (Skiera, Barrot e Becker, 2012) e, partindo disto, Katona, Zubcsek e Sarvary (2011) interpretam que a transmissão ou não de uma informação na rede se dá basicamente através da estrutura da rede em si (o grau a que os membros da rede têm inter-relações entre si - *clustering* - e o grau a que membros da rede estão localizados em pontos centrais das trocas de relação - *degree*) e pela relação social de influência entre os elementos participantes da difusão, apresentando estudos empíricos que começam a desvendar o papel da estrutura de uma rede social real, em larga escala, no contexto online.

Observando este panorama, Borgatti e Halgin (2011) sugerem que, ao passo que os estudiosos da administração e do marketing perceberam a iminência da necessidade de se conhecer a relação entre estrutura de rede social e a difusão e adoção de novos produtos, teorias de redes sociais foram resgatadas da sociologia para receber novos tratamentos nas disciplinas do marketing. Entre elas, estão a Teoria da Força dos Laços Fracos (SWT), de Granovetter (1973) e a Teoria do Buraco Estrutural (SHT), de Burt (2000).

Baseados nessas teorias, Katona, Zubcsek e Sarvary (2011) sugerem que pessoas localizadas em pontos de redes em que vizinhos estão altamente inter-relacionados devem ter características mais similares, pessoas em posições onde agem como ponto de ligação entre outros vizinhos devem ter maior controle sobre o fluxo de informação e, com isso, ter maior grau de influência e pessoas com maior quantidade de amigos devem ter mais chances de ser consideradas inovadoras e passíveis de imitação pelos membros da rede.

Estas sugestões ainda carecem de uma grande gama de comprovações empíricas, em diferentes contextos. O presente estudo pretende contribuir à academia estudando a relação entre a estrutura da rede social online e a o potencial de influência pessoal para a propagação de informações na rede social *Twitter*.

1.1 FORMULAÇÃO DO PROBLEMA DE PESQUISA

O campo de estudos das redes sociais como forma de propagar informações percebeu um “boom” de crescimento nos últimos anos. Diversas pesquisas foram elaboradas partindo do modelo de difusão de Bass (1969). Alguns pesquisadores dedicaram-se à estrutura da rede e outra parcela dedicou-se a estudar os atores. Quase todas essas pesquisas, senão todas, utilizaram-se de modelos de rede gerados por computador, deixando um chamado para comprovações empíricas, em redes sociais online reais.

Katona, Zubcsek, Miklos (2011) abriram uma frente a esse chamado ao estudar a propagação de uma nova rede social online européia, que replicaria as conexões que existem naturalmente off-line.

O estudo atual pretende contribuir estudando a propagação de informação em uma rede social real e estruturada, de abrangência mundial, respondendo: Qual o papel da estrutura da rede social online no processo de difusão de informação?

1.2 DEFINIÇÃO DOS OBJETIVOS DA PESQUISA

1.2.1 Objetivo Geral

O presente estudo pretende verificar o papel da estrutura da rede social online no processo de difusão de informação.

1.2.2 Objetivos Específicos

(1) Especificamente, pretendeu-se com a pesquisa verificar a relação entre quantidade de conexões, a proporção seguidores/seguidos de um membro da rede e a difusão de informação no ponto da rede social a que este membro está inserido, com base nas suposições do modelo de Bass (1969).

(2) Verificar a relação entre a quantidade de interconexões entre os vizinhos diretos de um membro de uma rede social online e a difusão de informação neste ponto específico, com base na teoria da Força dos Laços Fracos (SWT).

(3) Verificar a relação entre o grau a que um determinado membro de uma rede social online localiza-se em uma posição de menor caminho entre outros membros da rede e a difusão de informação neste ponto específico da rede, a partir da Teoria do Buraco Estrutural.

1.3 JUSTIFICATIVAS TEÓRICAS E PRÁTICAS

Apesar de ter seu início formal reconhecido na administração a partir do trabalho de Rogers (1976), as teorias de difusão de informações, sobre tudo no que diz respeito à difusão através do "boca a boca" em redes sociais, tem despertado grande interesse da academia ao passo que os meios de comunicação de massa evoluem para formatos mais digitais e sociais (Breiger, 2004). Kiss e Bichler (2008) apontam que, tanto para os acadêmicos do marketing, a fim de interpretar os fenômenos de difusão e adoção de novos produtos por parte de consumidores em rede social, quanto para os administradores de marketing, a fim de elaborar estratégias eficazes de lançamento e difusão de novos produtos, entender o papel de influência no ambiente online é ponto crucial.

Katona, Zubcsek e Miklos (2011) complementam esta observação ao oferecer o *insight* de que, ao estudar a estrutura da rede pode-se tanto identificar potenciais "atores" influenciadores e influenciados, como também obter informações precisas do comportamento do fluxo de informação em rede e apontam a necessidade de que se estude este processo, bem como a relação influenciador/influenciado em diferentes contextos de redes sociais, com diferentes aspectos.

A partir dessa construção, esse trabalho visa contribuir com a teoria do marketing ao:

- Relacionar as teorias de difusão de informação e as teorias de estrutura de rede social.

- Relacionar as teorias de influência com as teorias de rede social.
- Verificar, em ambiente real, a relação entre os elementos teóricos da estrutura de rede social e a difusão.
- Verificar, em ambiente real, a possibilidade de prever a difusão de informação a partir da estrutura da rede social.
- Verificar a possibilidade de analisar fenômenos de rede a partir de grandes bancos de dados.
- Construir parâmetros teóricos que expliquem o fenômeno da viralização.
- Construir parâmetros teóricos que possibilitem avaliar pontos de uma determinada rede social conforme suas características de difusão de informação.

E também contribuir com a prática do marketing ao:

- Criar possiblidades de prever a difusão de um produto, campanha ou mensagem a partir da estrutura da rede social online.
- Permitir uma melhor compreensão do uso do fenômeno da viralização para propagar mensagens em ambientes online.
- Permitir a possibilidade da identificação de personagens-chave para a divulgação de uma mensagem, a partir das características da estrutura de rede desses personagens.
- Verificar na prática a aplicabilidade das teorias sociais de rede e de difusão de informação no abiente de rede online.

2 REVISÃO DA LITERATURA

Este capítulo destina-se a descrever os principais temas que permeiam a compreensão do processo de difusão de informações dentro do contexto de redes sociais, mais especificamente o *Twitter*. Assim, os conceitos aqui tratados buscam embasar o quadro teórico proposto para o trabalho. Os temas discutidos são (1) a Adoção e difusão de novos produtos, sob a ótica da teoria de difusão de informações; (2) a estrutura de redes sociais para a difusão de informações, sob a ótica das teorias de estrutura de rede e (3) o funcionamento da rede social online *Twitter*.

2.1 ADOÇÃO E DIFUSÃO DE NOVOS PRODUTOS:

De acordo com Rogers (1976), os estudos da teoria da difusão têm início em dois pontos distintos: (1) o conceito de "difusionismo" encontrado em escolas de sociologia alemãs-austríacas e britânicas, em que assume-se que mudanças na sociedade se dão através da inclusão de inovações provenientes de outras sociedades e (2) o artigo do filósofo francês Tarde (1903) que pela primeira vez apresentou o gráfico de difusão como sendo uma curva em formato de "S", formada pela atuação de líderes de opinião e um sucessivo processo de imitação. Posteriormente, Ryan e Gross (1943) aplicaram o conceito ao estudar a difusão de novos tipos de sementes entre fazendeiros e abriram frente a mais de 4000 artigos e estudos em diversas áreas como inovação, agricultura, tecnologia, métodos de controle de fertilidade, economia e política (Wejnert, 2002).

Chandrasekaran e Tellis (2007, p.39), identificam que o termo tem sido utilizado diferentemente em dois grupos distintos de campos de estudo: (1) enquanto para os teóricos da economia e para a maioria dos teóricos das disciplinas não-marketing o termo difusão é definido como "a propagação de uma inovação através de grupos sociais ao longo do tempo", (2) no marketing e na comunicação, o fenômeno não é visto separado de seu condutor e, por isso, define-se como "a comunicação de uma inovação através da população". Visto que no marketing, aos olhos de Wright e Charlett (1995), a difusão é uma das teorias mais generalizadas,

Chandrasekaran e Tellis (2007, p.40) observam que parte dos estudiosos do marketing também é atraída para a definição utilizada na economia e, portanto, definem difusão como “a propagação de inovação através de mercados ao longo do tempo”.

No Marketing, o grande interesse dos pesquisadores por essa teoria se dá, segundo Wright e Charlett (1995), pelo ímpeto de reduzir as chances de falha de determinados produtos e pela necessidade de prever e acurar tomadas de decisões mercadológicas com relação a estratégias de lançamento e de difusão. Mahajan, Muller e Bass (1990) corroboram com essa ideia e identificam que enquanto o maior interesse dos pesquisadores de Comportamento do Consumidor tem sido, ao longo do tempo, avaliar a aplicabilidade das hipóteses desenvolvidas na área da difusão geral para os estudos de comportamento do consumidor, a literatura prática de administração de marketing tem buscado utilizar estas hipóteses para definir consumidores-alvo e desenvolver estratégias para atingir novos adotantes. Para tanto, Mahajan, Muller e Bass (1990) apontam que a pesquisa na administração e no marketing também contribuiu para os estudos da difusão desenvolvendo modelos analíticos para prever difusão de uma inovação em sistemas sociais e também desenvolvendo guias normativas de como uma inovação deve ser difundida.

Segundo Valente e Rogers (1995), Ryan e Gross (1943) abriram um campo de trabalho na teoria da difusão ao sugerir que, mais importante do que fatores estritamente econômicos, fatores sociais entre outros são responsáveis pela influência na adoção. Wright e Charlett (1995) reconhecem o modelo de difusão proposto por Rogers (1962, 1983) como um modelo amplamente aceito e difundido na academia, sobretudo no que diz respeito à abordagem teórica, ao passo que apontam melhorias e possibilidade de aplicação empírica provenientes do modelo proposto por Bass (1969).

Apesar de sugerir diversas limitações do modelo de Rogers (1962, 1983), Wright e Charlett (1995) apontam sua contribuição à teoria a partir do reconhecimento de que a curva de adoção de uma inovação (representando a frequência de consumidores adotando um produto ao longo do tempo) é normalmente distribuída em “forma de sino”, resultando, no caso de uma plotagem cumulativa do número de adotantes, em um padrão (sigmoide) em forma de “S”. Wright e Charlett (1995) também reconhecem a importância da definição dos quatro elementos necessários para a difusão de inovação.

Rogers (1962) explica como e por que uma ideia, comportamento ou objeto espalham-se por determinada população, afirmando que para que haja propagação de inovação, são necessários quatro elementos principais: a inovação, comunicação, tempo e um sistema social. Wejnert (2002) contribui à discussão apontando que outros determinantes da difusão foram apontados isoladamente em diversas áreas da ciência e ajuda agrupando estas variáveis em três componentes principais, sob a ótica sociológica: as características da inovação, do inovador e o contexto do ambiente.

2.2 A INOVAÇÃO

"Inovação", como conceito, é estudada em muitos contextos disciplinares, levando a uma multiplicidade de definições, incluindo psicológicas, sociológicas e econômicas (Brown e Ulijn, 2004). A inovação de Rogers (1983, p.11) é "uma ideia, prática ou objeto que é percebida como nova por um indivíduo ou pela unidade que a está adotando". Rogers (1976) observa que, mesmo este fato não sendo frequentemente reconhecido pelos estudiosos de difusão, ao longo da pesquisa de difusão da inovação, grande parte dos estudos voltou os olhos especificamente para a difusão de novos produtos, visto que uma boa parcela do comportamento de adoção de qualquer inovação está atrelada à compra de um novo produto.

Ao que tange a adoção de uma inovação, Rogers (2005) aponta cinco características facilitadoras do processo: (1) a vantagem relativa, ou seja, o grau em que a inovação é percebida como melhor do que a ideia que a antecede, podendo ser medida em termos econômicos, conveniência, prestígio social ou satisfação; (2) a compatibilidade, ou seja, o grau em que a inovação é consistente com os valores existentes, experiências e necessidades dos potenciais adotadores; (3) a complexidade, ou seja, o grau em que a inovação é percebida como difícil de adotar; (4) testabilidade, ou seja, o grau em que uma inovação pode ser testada antes de sua adoção completa; (5) observabilidade, ou seja, o grau em que os resultados da inovação podem ser vistos por outras pessoas do mesmo sistema social. Wejnert (2002) interpreta que a decisão de adoção a partir desses fatores reside em dois

eixos principais: consequências públicas X consequências privadas da decisão e custo para adoção (ou risco) X benefício da adoção.

2.2.1 O Tempo

O tempo é observado no processo de difusão, aos olhos de Rogers (1983), de três maneiras distintas. A primeira delas, para Rogers (2005), está envolvida na maneira pela qual a decisão de inovação é processada pelo potencial inovador. Esse processo se dá mentalmente ao passo que o indivíduo (ou a unidade de decisão) busca informações sobre a inovação a ser adotada a fim de reduzir riscos referentes a esta decisão Wright e Charlett (1985). Contemplando o momento em que a inovação é primeiramente percebida pela unidade decisória até a formação de uma atitude diante dela, que pode ser de adoção ou não, o processo se dá em 5 passos distintos, (1) a percepção da inovação, (2) o estabelecimento de uma atitude favorável ou não à inovação, (3) a tomada de decisão em relação à inovação, (4) o início da utilização da inovação (5) a avaliação dos resultados da adoção (Rogers, 2005).

A segunda maneira apontada por Rogers (1983) diz respeito ao potencial inovador da unidade de adoção e é visto como o grau em que o indivíduo é mais propenso a adotar a inovação antes de outros membros do sistema social ao qual ele está inserido. Rogers (1962) aponta cinco categorias de adotantes, baseado em seu potencial inovador: (1) os *Innovators*, representando os primeiros 2.5%, (2) os *Early Adopters*, os seguintes 13.5%, (3) a *Early Majority*, 34%, (4) *Late Majority*, 34% (5) *Laggards*, 16%. Estes “tipos ideais” levantados por Rogers (1983), como observam Wright e Charlett (1985), são baseados em características demográficas, de personalidade e socioeconômicas tendo cada um deles recebido um “perfil de consumidor” (Hawkins, Best e Coney, 1989). “No marketing, estas generalizações foram utilizadas como base para um guia prescritivo para acelerar o processo de difusão” (Wright e Charlett, 1995, p.35).

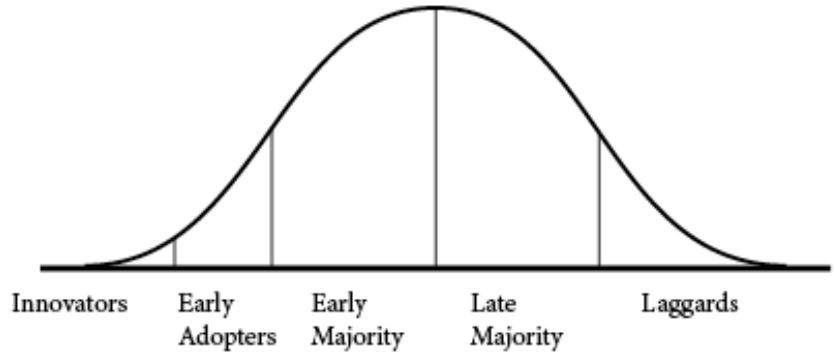


Figura 1 - Difusão de inovações.
Fonte: ROGERS, 1962.

Sob esta ótica, padrões de comunicação diferenciados para cada "perfil" foram propostos por Gatignon e Robertson (1985), mas Wright e Charlett, (1995) afirmam a existência de evidências empíricas demonstrando a inexistência de links consistentes ligando características de personalidade e a propensão a inovação.

Bass (1969, p.216), observando que esta discussão é largamente literária, sugere que "na formulação matemática da teoria apresentada aqui, devem-se agregar os grupos (2) a (5) e definí-los como imitadores". Esta sugestão vem a corroborar com Van den Bulte e Joshi (2006, p.1), que observam haver em mercados de consumo apenas dois tipos de consumidores, os *influenciadores*, "que estão em maior contato com novos desenvolvimentos" e os *imitadores*, "cujas próprias adoções não afetam os influenciadores".

Bass (1969, p.217) também diferencia imitadores de influenciadores através da influência de compra, sendo que "os inovadores, ao tempo de sua compra inicial, não são influenciados pela quantidade de pessoas que já adotaram o novo produto enquanto os imitadores são influenciados pelo número de compradores anteriores". Nesse contexto, Mahajan, Muller e Bass (1990) entendem que o modelo de Bass avalia que parte da influência de adoção reside na "imitação" ou "aprendizado" e, portanto está mais ligada à comunicação "boca a boca" enquanto outra parte não, e por isso pode estar ligada à influência dos meios de comunicação de massa.

A premissa do "modelo epidemiológico para a difusão de bens de consumo e outras inovações" (Wright e Charlett, 1995, p.36) testado e proposto por Bass (1969) reside na assunção de que "a probabilidade de compra a qualquer tempo é

relacionada linearmente ao número de compradores anteriores" (Bass, 1969, p. 226) e tem seu comportamento representado matematicamente por:

$$P(t) = p(0) + (q/m)Y(t)$$

P(t) é a probabilidade de compra em um tempo *t*, dado que o indivíduo não comprou a inovação antes, *p(0)* é a probabilidade inicial do indivíduo experimentar, que reflete sua tendência a inovação, *m* é o número total de compradores potenciais enquanto *q* é um parâmetro que mede o efeito de imitação, ou a taxa de difusão (Mahajan et al. 1990). *q/m* assume o papel de uma constante do efeito de interação social, dependente do tamanho do mercado e do efeito de influência interpessoal, que será magnificado pelo aumento no número total de pessoas que já compraram o produto, representado por *Y(t)* (Wright e Charlett, 1995).

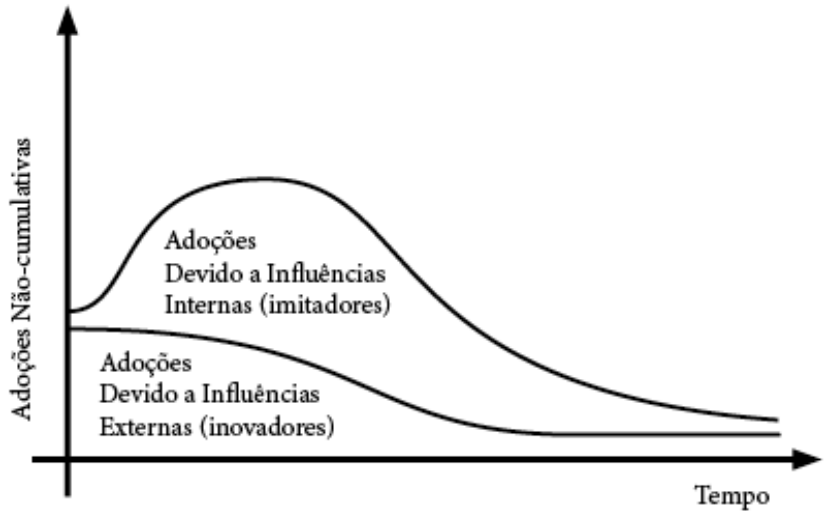


Figura 2 - O Modelo de Difusão de Bass - Adoções de acordo com influências externas e internas.
Fonte: MAHAJAN, MULLER E BASS, 1990.

Mahajan, Muller e Bass (1990) reconhecem a importância dos inovadores nos estágios iniciais da difusão de uma inovação, mas observam a queda dessa importância ao passo que a quantidade de adotantes aumenta e a influência interna ganha força.

A terceira maneira em que Rogers (1983) observa o tempo da difusão é a velocidade relativa com a qual uma inovação é adotada pelos membros do sistema social, a taxa de adoção. Wright e Charlett (1995) avaliam diversos estudos que testaram e validaram o modelo de Bass em diferentes tipos de difusão, comprovando a validade do modelo em períodos de difusão que variam de 5 a 50 anos (JEULAND, 1990; AKINOLA, 1986).

2.3 A COMUNICAÇÃO E O SISTEMA SOCIAL.

Já a comunicação, para Rogers (1983, p.17) “são os meios pelos quais a informação da inovação passa de um indivíduo para o outro”. Wright e Charlett (1995, p.32) reconhecem, dentro deste aspecto, tanto a validade dos canais de comunicação de massa quanto a comunicação interpessoal, apontando esta última, também conhecida como o “boca a boca”, como “o fator-chave para a velocidade e forma da curva de difusão”. Mahajan, Muller e Bass (1990) corroboram com essa visão ao afirmar que membros de um sistema social têm diferentes níveis de propensão para confiar na mídia de massa ou no “boca a boca”.

Rogers (1983, p.24) define o Sistema Social por “um grupo de unidades inter-relacionadas que estão engajadas em resolver problemas em conjunto para completar uma meta comum”. Desde os primeiros estudos de difusão conduzidos, segundo Valente (1995; 2006), observou-se que a natureza da difusão foi constatada como sendo um processo social envolvendo comunicação interpessoal entre indivíduos similares. Sob esta ótica, Kempe, Kleinberg e Tardos (2003) identificam que “uma rede social – o gráfico de relacionamento e interações entre um grupo de indivíduos – tem papel fundamental como meio para a difusão de informações, ideias e influência entre os membros de um grupo social”. Já Burt (2000, 2003, 2004) observa que a “forma” da rede de um indivíduo está relacionada com seu envolvimento social e com inovações.

2.3.1 Redes Sociais

O estudo de redes sociais, segundo Breiger (2004), tem origem em expressões metafóricas de vários pensadores e cientistas sociais (Marx, 1956; Durkheim, 1965; Wiese, 1941; Cooley, 1964). Estas expressões almejam descrever os resultados da relação entre grupos de indivíduos, representadas por linhas invisíveis cuja soma é a sociedade (Breiger, 2004). Bordieu (1986) vê nessas relações a agregação de recursos potenciais que, para Coleman (1990), definidas como capital social, facilitam ações individuais ou coletivas. Keenan e Shiri (2009) observam que a lente teórica multidisciplinar que avalia como a estrutura da rede afeta os usuários recebe o nome de Teoria das Redes Sociais ou *Social Network Analysis* (SNA).

Borgatti e Halgin (2011) apontam que o interesse na SNA por parte da academia cresceu exponencialmente ao longo do tempo em todas as áreas das ciências sociais e também na física, epidemiologia, biologia e administração. “Na pesquisa em administração, redes sociais têm sido usadas para entender a performance no trabalho, *turnove*r, promoção, inovação, criatividade, e comportamento anti-ético” (Borgatti e Halgin, 2011, p.1).

Partindo da definição da ciência social, Wasserman (1994) vê uma rede social como a representação de um grupo de indivíduos ou organizações (atores) interligados por laços didáticos, componentes de uma estrutura social (Krackhardt, 1998). Seguindo esta linha de pensamento, Martino e Spoto (2006, p. 53) afirmam que “qualquer tipo de agregação social pode ser representado em termos das unidades componentes dessa agregação e das relações entre estas unidades.

Corrobora com estas afirmações a definição encontrada em Landherr, Friedl e Heidemann (2012, p.6) que, baseados em Valente (1996) entendem o termo rede social como “um padrão de amizade, recomendação, comunicação ou apoio entre membros individuais ou grupos de membros em um sistema social”, cujos atores são unidos por metas, interesses ou necessidades comuns.

A gama de laços, ou potenciais relações, para Martino e Spoto (2006), é potencialmente infinita, visto que (1) cada laço pode carregar muitos significados diferentes (relações comerciais, amor, respeito, conexões físicas, links entre páginas de internet, avaliação entre um e outro, conhecimento, etc) e, como observam Katz

et al. (2004), (2) é comum que os atores compartilhem mais de um tipo de laço. Analisando a estrutura de laços e atores, diversas pesquisas contribuíram para a análise da difusão em rede social (Moody, 2009).

Granovetter (1973; 1982), ao avaliar a intensidade dos laços entre membros de uma rede observa que as redes sociais, no que diz respeito à intensidade, têm dois tipos de laços, (1) os laços fortes (tais como laços entre família e amigos), que geralmente envolvem grande grau de confiança e "são valiosos quando o indivíduo busca suporte sócio-emocional" (Katz et al, 2004, p.309) e (2) os laços fracos (tais como "conhecimento"), que envolvem menor grau de confiança mas "são valiosos quando o indivíduo busca informações diversas ou únicas (Katz et al, 2004, p.309).

Granovetter (1973) ilustra que as redes sociais são compostas de indivíduos densamente conectados através de laços fortes (formando clusters), que por sua vez são conectados entre si por laços fracos e esparsos (laços-pontes).

Esta teoria, como apontam Borgatti e Halgin (2011), recebeu o nome de *strength of weak ties (SWT)* [a força dos laços fracos, tradução aproximada] e é organizada a partir de duas premissas. A primeira afirma que "quanto maior a intensidade do laço entre dois indivíduos, maior a probabilidade que seus mundos sociais sejam similares" (Borgatti e Halgin, 2011, p.3). McPherson et al. (2001) observa que esta premissa parte do fato que os indivíduos tendem a ser *homófilos* e buscam ter laços fortes com indivíduos similares a eles. A segunda premissa é que laços "pontes" têm maior potencial de gerar boas ideias, já que ligam indivíduos com mundos sociais diferentes, "a ideia é que através de um laço-ponte, a pessoa pode ouvir coisas que não estejam circulando entre seus amigos mais próximos" (Borgatti e Halgin, 2011, p.4).

Burt (1992), avaliando a competitividade de informação entre membros da mesma estrutura de rede, aponta que indivíduos com maiores quantidades de ligações externas (aqueles ligados a pessoas de fora de sua rede de laços fortes) tendem a ter melhores performances em dadas circunstâncias, pois gozam de uma maior variedade de informações comparados àqueles com menos "laços não redundantes". Ao abordar a estrutura das conexões de redes, Baseado em Coleman (1988), Burt (2005), estabelece dois fenômenos sociais importantes. O primeiro interpreta que "quando dois indivíduos relacionados estão conectados à mesma terceira parte, a rede se torna melhor em transmitir informações" e isso faz com que as relações afetadas fiquem mais fortes (Katona, Zubcsek e Sarvary, 2011, p.7).

Burt (2005) batiza este fenômeno de *network closure* [fechamento de rede, tradução do autor] e complementa que este tipo de estrutura aumenta a confiança da informação, pois cria caminhos redundantes.

O segundo fenômeno apresentado por Burt (2005) diz respeito à importância social dos indivíduos cujos laços-ponte interconectam clusters. Katona, Zubcsek e Sarvary (2011) observam que estes indivíduos podem influenciar todo o seu *cluster*, pois têm poder de controle e "curadoria" da informação vinda de outros grupos. O fenômeno recebe o nome de *brokerage* [corretagem, tradução do autor] e a teoria é batizada de *structural hole (SH)* [teoria do buraco estrutural] (Burt, 1992). Kunst e Kratzer (2007, p.38) apontam que "atores ligados diretamente a buracos estruturais tem acesso a diversas informações antes dos outros, o que lhes dá melhor acesso a inovações e maior capacidade de distribuir boas ideias".

Como observado por Kilduff (2010), apesar de assumir uma visão mais estratégica e instrumental, a SH é muito similar à SWT em suas conclusões. Borgatti e Halgin (2011, p.5) corroboram com esta percepção e interpretam que "a razão pela qual os laços fracos são úteis é justamente porque eles fazem pontes entre diferentes clusters", ou seja, "é seu buraco estrutural que os torna avantajados".

2.4 ESTRUTURA DE REDE E DIFUSÃO: HIPÓTESES DA PESQUISA

Moody (2009) aponta que diversos estudiosos já reconheceram a importância da estrutura da rede para o fluxo de informação (Coleman, Katz e Menzel, 1966; Valente, 2001, 1995; Rogers, 1962). Liu, Madhavan e Sudharshan (2005, p. 242) explicam que "a ideia principal na tradição dos estudos de rede é que a estrutura social em si influencia a transmissão de novas ideias e práticas moldando os padrões de interação na rede". Partindo desta percepção, Scott (1991) também enfatiza que, aos olhos da SNA, para entender a difusão de informação em rede, a estrutura social da rede em si é mais importante do que os próprios usuários individuais.

Compreendendo a relação inovador/imitador da teoria de difusão, observando o potencial risco a que um indivíduo se submete ao adotar uma novidade e as chances desse risco ser minimizado pela imitação de um comportamento já existente na rede, Liu, Madhavan e Sudharshan (2005, p.243)

explicam que “a intuição fundamental da teoria de difusão em rede é que os padrões estruturais da rede determinam quem um dado ator escolherá como modelo”.

Moody (2009) reúne três características identificadas teoricamente como formadoras de difusão. A (1) *conectividade*, refere-se ao “sistema de caminhos formado pela concatenação das redes locais”, o (2) *clustering* [agrupamento] refere-se à “probabilidade que um caminho iniciando em um nodo retorne ao mesmo nodo inicial” e o (3) *degree features* [grau de centralidade], que se refere “à quantidade e ao padrão dos contatos diretos de um mesmo nodo” (Moody 2009, p.5). Landherr, Friedl e Heidermann (2010) corroboram com a percepção da importância desses três elementos e abordam, como medida de centralidade influenciadora da difusão, também o fenômeno (4) *betweeness centrality*, referente à presença de um determinado nodo da rede nos menores caminhos possíveis entre outros nodos.

2.4.1 Conectividade

O conceito de conectividade engloba (1) a definição de componentes (como sendo o maior conjunto de nodos que tem pelo menos um caminho conectando cada par de atores), como se fossem peças completas de rede e (2) o tamanho do caminho mais curto que a informação deve percorrer para passar de dois pontos quaisquer em uma rede, a *path distance* (Newman, 2001). De acordo com Watts e Strogatz (1998), quanto maior a *path distance*, menores serão as chances de difusão. Similar ao conceito de redundância de Burt (1992), caminhos independentes ligando os mesmos dois pontos aumentam as probabilidades de difusão (Moody e White, 2003).

2.4.2 Clustering [ou aglomeração, por tradução livre]

Para Katona, Zubcsek e Sarvary (2011), o efeito de *clustering* é relevante para a adoção ao passo que há grande relação de influência entre membros de grupos altamente interconectados. Sob esta ótica, espera-se que **H_1:** Haja uma

relação significativa positiva entre o grau de *clustering* de um determinado ponto de uma rede e a presença de contágio da mensagem nesse ponto.

Em contrapartida, ao analisar os clusters, Moody (2009) ressalva que nesse tipo de estrutura há a possibilidade de um efeito negativo à eficiência da difusão, ao passo que essa gama de interconexões podem fazer com que o objeto da difusão trafegue internamente pelo cluster sem conseguir avançar de maneira eficiente para outros pontos da rede, à medida que os caminhos voltam a si mesmos.

Observado por este ponto de vista, o efeito *clustering*, apesar de ter potencial efeito negativo na difusão a longo prazo, tem efeito positivo na probabilidade de adoção em redes em que já houve adoção. Liu, Madhavan e Sudharshan (2005) identificam, portanto, que atores envolvidos em redes de "alta densidade", aquelas com maior grau de *clustering*, têm maior propensão à imitação pois (1) sofrem de maiores pressões comportamentais e (2) estão expostos a maior redundância de informação.

Partindo desta construção teórica, espera-se que **H_2:** haja correlação significativa positiva entre o grau de clustering de vizinhos adotantes de um determinado ponto de uma rede e a eficiência de contágio relativa nesse ponto.

2.4.3 *Degree* ou [grau de conectividade, por tradução livre]

Construindo sobre a sugestão de Granovetter (1973) e Valente (2005), em que um membro da rede pode ter maior grau de influência sobre o comportamento de outros membros quando o outro tem, no total, uma quantidade menor de amigos, Katona, Zubcsek e Sarvary (2011), concluem que, intuitivamente, se o adotante tem mais amigos já utilizando um produto ou serviço, há maiores chances de que ele venha a também adotar. Moody (2009) apresenta este fenômeno como *degree* e entende que, quanto maior o *degree* (ou grau de conectividade) de um membro da rede, mais potencialmente influente ele se torna. Por isso, espera-se que **H_3:** a presença de contágio da mensagem em um ponto da rede tenha correlação significativa positiva com a proporção entre a quantidade de conexões deste ponto da rede e a média de conexões de seu vizinho.

Corroborando com o que é proposto na teoria SWT, Lin (1999) afirma que redes de contatos maiores oferecem maior facilidade para receber informação e, portanto, espera-se que os elementos altamente conectados em uma rede, batizados de "hubs" (Landherr, Friedl e Heidermann, 2010), sejam lideres de opinião (Katona, Zubcsek e Sarvary, 2011). Sob esta ótica, espera-se que **H_4:** Quanto maior a quantidade de conexões de acesso a informação de um ponto da rede, maior será a chance de haver presença de contágio da mensagem nesse ponto.

Goldenber et al (2006) identificam condições em que as características da conectividade social dos membros de uma rede é mais significante do que expertise no que se diz respeito à influência de outros membros. Partindo deste aspecto espera-se que **H_5:** Quanto maior a quantidade de conexões de envio de informação de um ponto da rede, maior será a chance de haver presença de contágio da mensagem nesse ponto.

French e Raven (1960) e Stephen et al (2010), no entanto, questionam a abrangência dessa conclusão ao propor, sob a ideia que todos os indivíduos têm um poder de influência limitado, que o maior poder de influência dos hubs restringe-se aos seus laços mais próximos.

Liu, Madhavan e Sudharshan (2005) identificam que os atores em posições altamente centrais na rede, graças ao grande acesso a informação citado em Lin (1999), encontram-se em boa posição para inovar (pois têm recursos de sobra, o que fomenta a experimentação) e têm predisposição negativa a copiar membros da rede em posições de menor centralidade (apenas inovando ou imitando parceiros com ainda maiores centralidades). Assim, espera-se que **H_6:** Quanto maior a proporção entre a quantidade de conexões de envio de informações e a quantidade de conexões de recebimento de informações de um ponto da rede, maior será a chance de haver presença de contágio da mensagem nesse ponto.

Katona, Zubcsek e Sarvary (2011) ainda observam que quanto maiores as redes de contatos de determinados nodos com o mesmo grau de *clustering*, maior será a influência desses nodos diante de seus vizinhos e apontam uma grande relação entre a interação entre o *degree* e o *clustering* na difusão de uma mensagem.

2.4.4 *Betweeness* [ou Centralidade-meio, tradução de Kirschbaum (1996)]

Para Katona, Zubcsek e Sarvary (2011), o *betweeness centrality* é construído por Freeman (1977, 1979) partindo do entendimento que se, para cada caminho entre dois pontos (A, C) da rede, o menor caminhos possível envolve a passagem por B, então este par de nodos contribuí para a *betweeness centrality* de B. Landherr, Friedl e Heidermann (2010) observam que, nesse caso, B pode exercer controle no fluxo de informação entre A e C, já que a comunicação entre estes dois pontos dependem da interveniência de B. Liu, Madhavan e Sudharshan (2005) também observam que B também tem vantagem no fluxo de informação pois pode receber informações diretas tanto de A quanto de C, com redundância mínima e ressalvam que se A e C também conectam-se entre um e outro, mesmo que indiretamente, o poder exercido por B passa a reduzir. Construindo sobre esta observação, membros de redes que não representam buracos estruturais (com baixa *betweeness centrality*), também chamadas de *constraints*, têm menor propensão a inovar (Liu, Madhavan e Sudharshan, 2005).

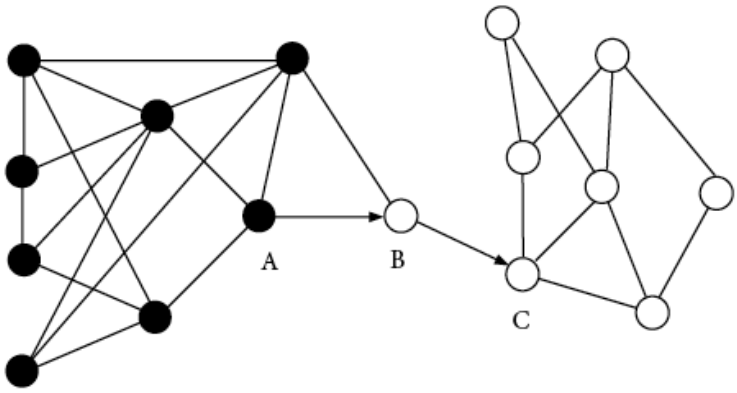


Figura 3 - Representação de Betweeness Centrality de "B" na relação "A" e "C".

Esta discussão, segundo Landherr, Friedl e Heidermann (2010) tem há muitos anos tomado posição de destaque na SNA, mesmo sendo limitada a redes sociais indiretas ou fictícias. Nos últimos anos, contudo, a evolução dos meios de comunicação digital permitiram aos estudiosos de redes sociais a possibilidade de

analisar dados reais, de redes reais, em contextos reais (Katona, Zubcsek e Sarvary, 2011).

2.4.5 Eigenvector Centrality [ou centralidade do vetor característico]

Boniacich (1972) desenvolveu uma forma de avaliar a centralidade de um membro de uma rede não apenas a partir da proporção entre o número de seguidores dele com relação ao número de seguidores dos outros usuários, mas também pelo grau de centralidade dos usuários conectados a ele. Esta medida é conhecida como Eigenvector Centrality e, para Boniacich (2007), oferece grandes vantagens no cálculo do grau de centralidade de membros de redes muito complexas, se comparado aos outros indicadores de centralidade (Betweeness, Degree, etc), que foram elaborados teoricamente, a partir da analise de redes sociais menos complexas.

H_7: quanto maior a eigencevtor centrality de um ponto, maior será a chance de haver presença de contágio da mensagem nesse ponto.

2.5 CONTEXTO DAS REDES SOCIAIS NA INTERNET.

De acordo com Libai et al. (2010), a última década observou desenvolvimentos significativos no entendimento dos antecedentes e das consequências das relações C2C (Customer to Customer). Essa evolução se dá (1) pela maior conectividade entre os consumidores, proporcionada por novas tecnologias de comunicação, (2) pelo fato de os gestores e acadêmicos terem à sua mão informações que historicamente foram impossíveis ou muito difíceis de obter e (3) pela percepção gerencial da necessidade do engajamento entre marcas e consumidores (Libai et al, 2010).

“Historicamente, o estudo de redes sociais foi feito baseado em um pequeno número de bases de dados repetitivamente analisadas” (Libai et al, 2010, p.267). Hill, Provost e Volinsky, 2006), no entanto, apontam para uma nova infinidade de

dados, de múltiplas fontes, de redes sociais de grande escala disponíveis a partir das redes sociais online. Landherr, Friedl e Heidemann (2010) comentam o grande crescimento das plataformas de rede social online, exemplificada pelo *Facebook.com*, que supera o buscador *google.com* em visitas diárias e a representativa taxa de uso dessas plataformas, exemplificados pelo dado que 66% dos usuários de internet participam de comunidades sociais online ao menos uma vez por mês. Mislove et al. (2007), reconhecendo que os sites de relacionamentos redes sociais online (como Orkut, Flickr, Facebook e Youtube) estão entre os mais populares da internet, chama atenção para a necessidade de estudar as características de Redes Sociais Online (OSN), baseadas em Sites de Social Networking (SNS).

Sites de Redes Sociais, para Keenan e Shiri (2009, p.439), são tecnologias para socialização online através de “websites que encorajam a interação social através de contas de usuários baseadas em perfis”. Diferente da web, que é organizada por conteúdo, Mislove et al. (2007) observam que os sites de redes sociais online são organizados por usuário. Nesse contexto, os usuários participantes entram na rede preenchendo um perfil de usuário (que pode envolver pseudônimos) e criam links para outros usuários a fim de manter relacionamentos sociais e linhas de comunicação com eles (Mislove et al, 2007). As redes sociais também são utilizadas, de acordo com Gneiser et al. (2010, p.3) “para encontrar e trafegar por conteúdo aprovado/indicado por outros usuários”.

3 METODOLOGIA

Esse capítulo pretende descrever o delineamento metodológico da presente pesquisa, e visa apresentar os passos adotados para a realização da pesquisa, e resumir o esforço do pesquisador para obter os dados de maneira a atingir os objetivos da pesquisa com a maior eficácia possível. (RICHARDSON, 1999, p. 138). Para tal, o terceiro capítulo está dividido em duas partes. A primeira, entitulada “Especificação do Problema” contempla a apresentação do modelo de estudo, as hipóteses da pesquisa e as definições constitutivas e operacionais das variáveis. Na segunda parte, são tratadas a delimitação e o design da pesquisa, abordando a população, critérios para a composição da amostra, as fontes de dados, a estratégia de coleta, o tratamento dos dados e as limitações do estudo.

3.1 ESPECIFICAÇÃO DO PROBLEMA

Nesta parte do trabalho, o modelo proposto é apresentado e as hipóteses da pesquisa são descritas. Para tal, também são apresentadas as definições constitutivas e operacionais das variáveis.

Propondo investigar a relação entre dois temas específicos, a Difusão de Informação e a Análise de Redes Sociais, o questionamento-guia dessa investigação se deu pelo ímpeto de descobrir o papel da estrutura da rede social online no processo de difusão de informação.

3.1.1 Perguntas de Pesquisa

A partir do problema destacado e dos objetivos apresentados anteriormente, o presente estudo buscará estudar as hipóteses apresentadas no referencial teórico orientado pelas seguintes perguntas de pesquisa:

- Qual a relação entre quantidade de conexões, a proporção seguidores/seguidos de um membro da rede e a difusão de informação neste ponto específico da rede social?
- Qual a relação entre a quantidade de interconexões entre os vizinhos diretos de um membro de uma rede social online e a difusão de informação neste ponto específico?
- Qual a relação entre o grau a que um determinado membro de uma rede social online localiza-se no menor caminho entre outros membros da rede e a difusão de informação neste ponto específico da rede?

3.1.2 Definição constitutiva e operacional das variáveis

A definição constitutiva de uma categoria de análise compreende a sua descrição conceitual, com base na teoria apresentada, e pretende esclarecer a teoria e conceitos a ser utilizados para interpretar um determinado fenômeno. Triviños (2010) justifica que, por muitas vezes tratar de conceitos abstratos, faz-se necessário a apresentação da aplicação prática de cada conceito em relação à análise que será feita no estudo.

Para tanto, esta parte do trabalho destina-se a apresentar os tópicos que dizem respeito às definições constitutivas (DC) e as definições operacionais (DO) das categorias de análise deste estudo.

Para responder especificamente às perguntas da pesquisa, atendendo aos objetivos, faz-se necessário aprimorar a definição das variáveis a ser estudadas.

3.1.2.1 Conexões

DC– Granovetter (1973; 1982), observa que a partir dos laços formados nas redes sociais, podem haver diversos tipos e intensidades de relações e sentidos de transferência de informação.

DO– A variável *Conexões-E* é constituída, teoricamente, como a contagem de todos os elementos para os quais um determinado elemento da rede envia suas mensagens. No *twitter*, pode-se optar por seguir ou não um usuário. Todos os usuários que optaram por seguir este usuário compõe o número de seguidores dele. Para a variável *Conexões-E* da presente pesquisa, atribuiremos a quantidade de usuários do *twitter* que optaram por seguir o usuário em questão (S) - os perfis para os quais ele envia mensagens.

DO – A variável *Conexões-R* é constituída, teoricamente, como a contagem de todos os elementos dos quais um determinado elemento da rede recebe mensagens. Todos os perfis que um determinado usuário optou por seguir compõe o número de seguidores desse usuário. Para a variável *Conexões-R* da presente pesquisa, atribuiremos a quantidade de perfis do *twitter* que o usuário em questão optou por seguir (Z) - os perfis dos quais ele recebe a mensagem.

3.1.2.2 Popularidade Relativa

DC – Goldenber et al (2006) observam que as características da conectividade social e do fluxo de informação em um determinado nodo da rede, tais como a quantidade de informação que chega e que sai do nodo, refletem o grau de influência do nodo. Landherr, Friedl e Heidermann, 2010 entendem que a popularidade de um ponto da rede está ligada à relação entre a quantidade de pessoas que recebem mensagens desse nodo e a quantidade de pessoas que enviam informação para ele.

DO – Na rede social *Twitter*, para esta pesquisa, pretendemos adotar o conceito de popularidade relativa (P) como uma variável que mede a proporção entre a quantidade de “seguidores” (followers) de um determinado ponto da rede (S) e a quantidade de pessoas que o mesmo ponto segue espontaneamente (Following, Z).

$$P = \frac{S}{Z}$$

3.1.2.3 Clustering

DC–Para Moody (2009), *clustering* refere-se à quantidade de interconexões entre um grupo de nodos de uma rede e a probabilidade que a comunicação, iniciando em um ponto específico, retorne a este mesmo ponto pelos caminhos da rede.Katona, Zubbcsek e Sarvary (2011) apresentam o *clustering* como a variável que mede a densidade das interconexões entre usuários. Batizamos de Clustering-N a medida do *clustering* no nodo específico avaliado no estudo e de Clustering-M a medida do *clustering* do "contágio" da mensagem a ser avaliada no nodo específico.

DO – No presente estudo, a variável Clustering-N(Cn) medirá o clustering do usuário do *twitter* a ser estudado e será obtida através do resultado da proporção entre o número de conexões existentes entre todos os usuários ligados a um determinado usuário do *twitter*(Cn) e ele mesmo e o número de conexões possíveis entre estes mesmos usuários (Yn).

$$Cn = \frac{yn}{Yn}$$

DO – No presente estudo, a variável Clustering-M (Cm) medirá o clustering da mensagem entre a rede direta do usuário do *twitter* a ser estudado e será obtida através do resultado da proporção entre o número de conexões existentes entre todos os nodos ligados a um determinado nodo que já tenham disseminado a mensagem (Ym) do estudo e ele mesmo e o número de conexões possíveis entre estes mesmos nodos (Ym).

$$Cm = \frac{ym}{Ym}$$

3.1.2.4 Degree

DC–Valente (2005) entende que um membro de determinada rede tem maior grau de influência sobre seus vizinhos caso estes vizinhos tenham, em média,

uma quantidade menor de amigos. Esta relação é batizada por Granovetter (1973) como degree centrality).

DO – Operacionalmente, definimos o Degree (X) para este estudo como sendo a proporção entre a quantidade de seguidores de um determinado usuário (s)de *twitter* e a média de seguidores de seus seguidores ($\overline{s_s}$).

$$X = \frac{s}{\overline{s_s}}$$

3.1.2.5 Eigenvector Centrality

DC – Boniacich (2007), define a Eigenvector Centrality como o score relativo a um nó da rede comparado ao score dos outros nós da rede. Os nós conectados a outros nós com maiores scores têm maiores scores.

DO – Neste estudo, foi medido o Eigenvector Centrality *(Ec)*a partir da fórmula abaixo, onde *Eci* é o Eigenvector Centrality do nó *i* e *A=(aij)* sendo a matriz de adjacência da rede. Portanto *aij = 1* se o nó *i* está ligado ao nó *j*, e caso contrário *aij = 0.* Generalizando, as entradas em *A* podem ser números reais representando as forças de conexão, como em uma matriz estocástica.

Para o nó *i*, o score de centralidade eigenvector será proporcional à soma dos *scores* de todos os nós aos quais ele está conectado.

$$Eci = \frac{1}{\lambda}\sum_{j=1}^{N} aij.Ecij$$

3.1.2.6 Contágio

DC – Valente (1996) atenta que diferentes pontos de uma rede têm diferentes graus de eficiência ao transmitir uma inovação, no que tange a adoção por parte dos vizinhos e utiliza o termo contágio para definir a eficiente transmissão e adoção de uma inovação através dos laços da rede.

DO–Neste estudo, foi medido o contágio (*D*) como sendo o número de vizinhos deste mesmo usuário que adotaram o uso de uma #hashtag após a emissão por parte do usuário (*k*).

$$D = k$$

DO – Também foi medido o contágio relativo(*Drel*) de cada membro da rede a partir da proporção entre o número de vizinhos deste mesmo usuário que adotaram o uso de uma #hashtag após a emissão por parte do usuário (*k*) e a quantidade total de seguidores deste usuário (*s*).

$$Drel = \frac{k}{s}$$

DO – Para analisar a relação causal do contágio, também foi criada a variável contágio condição (*Dbin*) considerando a presença de adoção da variável por parte dos vizinhos (1) ou ausência (0).

$$Dbin = 0 \; ou \; 1$$

3.1.3 Modelo

Com base nas hipóteses previamente apresentadas no referencial teórico e nas variáveis definidas e constituídas para o estudo, o modelo proposto espera contribuir para a compreensão da relação entre as características estruturais das redes sociais e a difusão da informação ou contágio, conforme mostrado na figura 4.

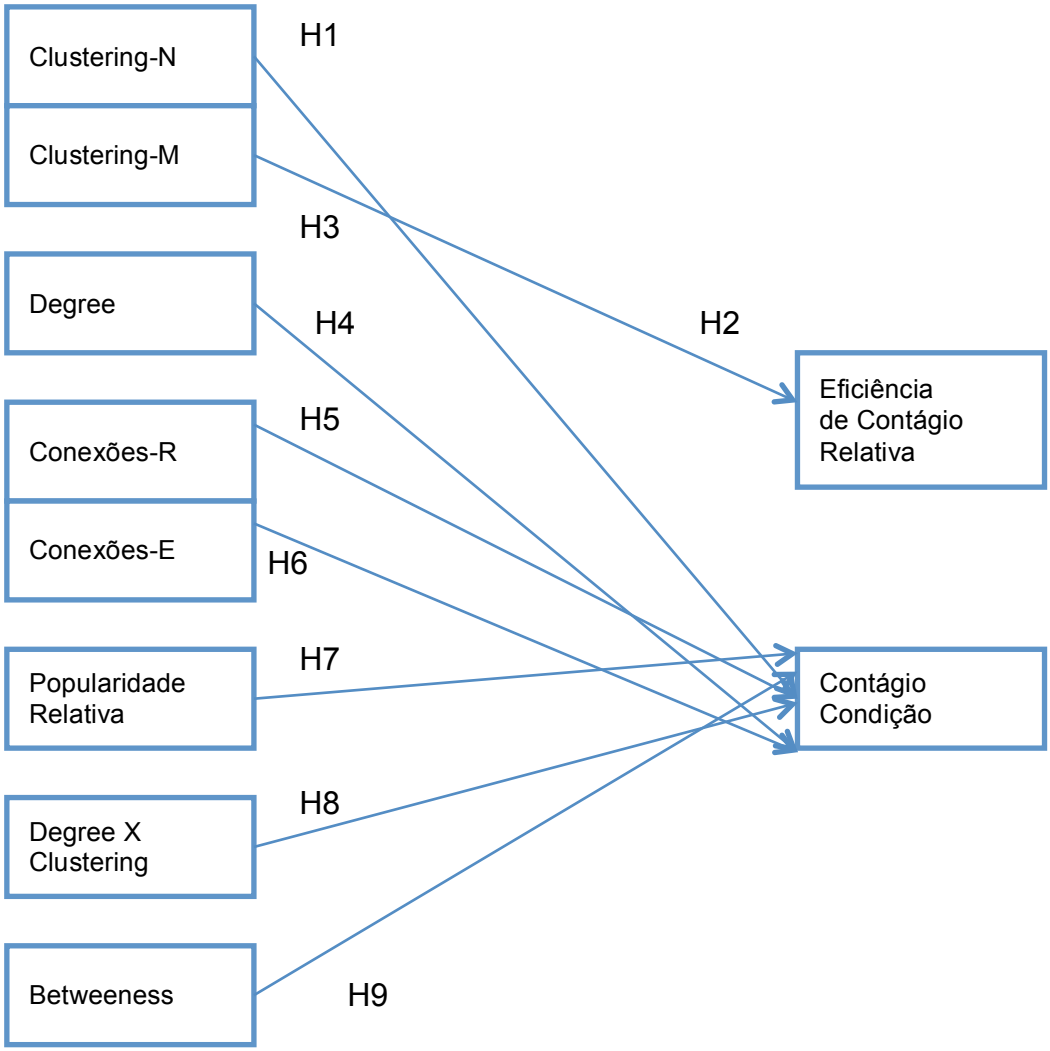


Figura 4 – Modelo proposto.

3.2 DELIMITAÇÃO DA PESQUISA

Por incluir a definição do método utilizado na presente pesquisa e contemplar a gama das atividades realizadas a fim de guiar a coleta e análise dos dados, para garantir a significância para o objetivo da pesquisa e otimizar o processo, esta etapa do trabalho apresenta como a pesquisa foi organizada a fim de otimizar os recursos e tempo necessários para alcançar os objetivos propostos. (SELLTIZ et al, 1967). Estão contemplados neste item o delineamento da pesquisa, onde estão detalhadas as etapas percorridas para que o desenvolvimento do projeto, e o procedimento utilizado para obtenção da amostra.

3.2.1 Delineamento da Pesquisa

O estudo caracterizou-se como descritivo, que segundo Triviños (1987), proporciona conhecer as características da comunidade ou grupo objeto do estudo, em termos contextuais de problemas, cultura, valores, distribuição nas categorias de análise e outros. O estudo descritivo também serve para registrar, analisar e correlacionar fatos ou fenômenos (variáveis), para descobrir a freqüência do fenômeno, relações e conexões com outros fenômenos e características.

Segundo Pinsoneault (2001), a pesquisa em questão foi explicativa, já que testou um conjunto de teorias a partir de relação entre variáveis.

Por ter intenção de testar a relação entre as variáveis de estrutura de rede e a difusão diante das teorias de difusão de informações e das teorias da SNA o presente estudo, de acordo com Creswell (2010), caracterizou-se pela abordagem quantitativa.

O tipo de investigação foi um levantamento por amostragem para obtenção de dados da disseminação de informação no *twitter*. Os dados secundários sobre os elementos da amostra foram obtidos de relatórios disponíveis na plataforma do *twitter*.

Quanto à estratégia, o estudo foi do tipo pesquisa de levantamento, conforme Pinsouneault (2001), e visou proporcionar uma descrição da estrutura da rede *Twitter* e da difusão de informações nesse contexto. Optou-se pelo levantamento de informações sobre uma população a partir de dados amostrais. No presente trabalho, por utilizar informações de opiniões ou percepções de usuários, não se aplicam as deficiências da estratégia survey apontadas por Malhotra (2001), tais como a perda de dados como sensações e crenças, ou a dificuldade de descrever situações, mas são válidas as vantagens do método, como a redução da variabilidade dos resultados, a aplicação simples e a confiabilidade dos dados.

Visto que o objeto principal do trabalho foi verificar a relação entre variáveis em um espaço específico de tempo, o levantamento foi longitudinal, aos olhos de Creswell (2010), a partir do corte temporal realizado - compreendeu dados de junho e julho de 2012, assume, a característica de ex-post-facto não-experimental, uma vez que nenhuma das variáveis independentes foi manipulada pelo pesquisador (KERLINGER, 1980), mesmo que tenham sido usados no presente estudo de

maneira semelhante ao levantamento cross-sectional, visto que – para o fim desse estudo – não se levou em consideração as diferenças temporais dentro do corte longitudinal.

O nível de análise é o de grupo social e a unidade de análise ou o sujeito da pesquisa foram as pessoas que recebem e retransmitem mensagens no *twitter* e, com os dados sobre a transmissão de mensagens de um determinado tema e das características das pessoas.

As hipóteses foram testadas com a utilização de análises multivariada de dados, especificamente regressões logísticas e regressão linear múltipla. Para tanto, a presente pesquisa teve as seguintes etapas:

- Revisão da literatura – conforme apresentado anteriormente.
- Desenvolvimento do modelo teórico de análise – conforme apresentado.
- Identificação da população e definição da amostra – a partir de pesquisa de assuntos que tenham volume de difusão suficiente para obtenção de uma amostra significativa.
- Construção do instrumento de pesquisa – desenvolver programa computacional para coletar os dados da plataforma *twitter*.
- Validação do instrumento de pesquisa – obtenção dos dados em uma pequena amostra para correção e ajustes.
- Coleta e processamento de dados – obtenção dos dados durante período necessário para completar a amostra necessária e a construção do banco de dados.
- Tratamento de dados – análise de dados perdidos, outliers e normalidade das variáveis.
- Análise dos resultados – análise das relações entre as variáveis para teste das hipóteses.
- Conclusões.

3.2.2 População e amostra

Em linha com a visão de Malhotra (2001), que define ser a população todos os elementos de características semelhantes, capazes de responder à investigação, a população do estudo foi todos os usuários da plataforma *Twitter*.com.

O EBIZMBA (2012), que ranqueia os sites de rede sociais mais populares aponta o *facebook.com* e o *twitter.*com os primeiros em popularidade, com 750.000.000 e 250.000.000 usuários mensais estimados respectivamente. A escolha do *twitter* deve-se à relevância dessa plataforma, do formato da rede adequar-se a análise das relações das variáveis do modelo de análise e da possibilidade do acesso aos dados.

Os detalhes da plataforma *Twitter.com* e dos seus usuários são apresentados no capítulo 5 e o detalhamento do acesso aos dados, suas limitações e vantagens estão descritos a seguir.

3.2.2.1 O acesso aos dados do *Twitter*

Segundo *TWITTER* (2012), a leitura de tweets e perfis de usuários (que não optaram por deixar seu perfil e *tweets* indisponíveis para pessoas de fora de sua rede de contatos) pode ser feita por qualquer pessoa, enquanto a possibilidade de postar mensagens (pelo computador, por SMS ou por dispositivos móveis) só é oferecida a usuários registrados.

3.2.2.2 Interface de comunicação com outros aplicativos

O *TWITTER* DEV BLOG (2013), ao analisar o "ecossistema do *twitter*", afirma que além dos usuários "comuns" da plataforma, também existem uma grande gama de empresas e desenvolvedores individuais construindo aplicativos que interagem com o *Twitter* para realizar analises ou criar engajamento (figura 6). Esses

aplicativos acabam muitas vezes utilizando o espaço destinado aos usuários comuns.

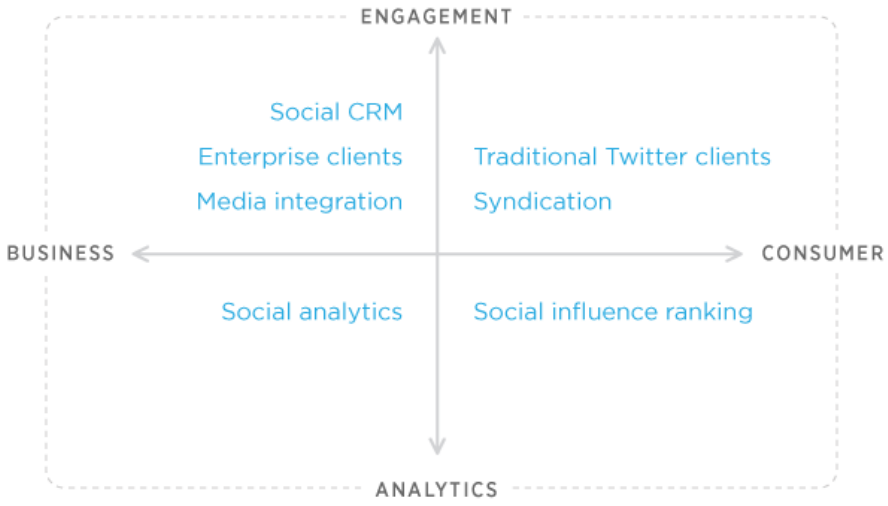


Figura 5 - Quadrantes de utilização do ecossistema *Twitter.*
Fonte: Twitter Dev Blog, 2013.

Para limitar grandes volumes de transferências de dados realizadas por aplicativos externos no quadrante “Traditional *Twitter* clientes Syndication”, que podem atrapalhar o desempenho da plataforma *Twitter* para o usuário final, e estimular as interações nos quadrantes “Social influence ranking”, “Social analytics” e “Enterprise clientes Media integration”, o *TWITTER* DEV BLOG (2013) explica que a *Twitter* API (Application Programing Interface – nome dado à parte do sistema da plataforma *Twitter* que interage com os aplicativos externos) permite a comunicação entre aplicativos externos ao *Twitter* e o banco de dados da plataforma, mas limita o volume de dados a serem transferidos.

3.2.2.3 Limitações da *Twitter* API

O *TWITTER* DEV BLOG (2013) explica que a versão 1.1 da *Twitter* API, requer, na interação com aplicativos externos, as seguintes ações, disponibilidades e configurações:

Autenticação: Requisições não identificadas à *Twitter* API não são aceitas. Para comunicar-se com a interface, o sistema deve estar autenticado com alguma identidade de usuário cadastrado no *Twitter*. Isso permite que o *Twitter* tenha maior controle e entendimento sobre quais tipos de aplicações estão acessando o sistema.

Limitação de Taxa de Dados por Usuário Autenticado: Cada ponto da rede *Twitter* (usuário autenticado) pode fazer um máximo de 60 requisições à API em uma hora (15 a cada 15 minutos). Essa limitação protege a interface de requisições abusivas.

3.2.2.4 Vantagens do Twitter em relação a outras redes sociais

Levando em consideração o formato, estrutura e disponibilidade da rede, a utilização do *Twitter* como plataforma de rede a ser analisada apresenta uma série de vantagens para o estudo proposto, como:

Relevância:

- Ser uma rede social altamente relevante em volume de usuários.
- Ser uma rede social de abrangência mundial.
- Ter grande fluxo de comunicação entre os usuários.

Transparência:

- Disponibilizar publicamente maior quantidade de informações, comparada às outras redes.

Estrutura Operacional:

- Utilizar o sistema de *#hashtags* (palavras-chave que simbolizam o assunto, montadas no formato “#” + “palavra”) para organizar tópicos de discussão.
- Oferecer aos visitantes do site os tópicos (frases, palavras ou *#hashtags*) mais frequentes do dia através da ferramenta *Twitter Trending Topics*.

Estruturação particular da Rede:

- Oferecer aos membros da rede a possibilidade da criação de laços não-recíprocos (um membro pode ou não optar por seguir - criar laço - com outro, independentemente da decisão ou autorização do último).

3.2.2.5 Definição Geral da População

Sendo assim, a população do estudo proposto são os perfis dos usuários de site *twitter.com* que já utilizaram sua conta para enviar mensagens, já fazem parte da rede através do ato de *seguir* e *ser seguido* por outros membros e se engajaram na difusão de *hashtags* apontadas pelo próprio *twitter*.com como os *Trending Topics* mundiais das nove semanas entre 01/julho e 01/setembro de 2012.

3.2.2.6 Processo de decisão pela amostra

Diante das limitações da *Twitter* API apresentadas nos capítulos anteriores, buscando atender os objetivos da pesquisa, a amostra das *hashtags* foi por conveniência, intencional e não-probabilística de acordo com os temas mais relevantes discutidos no *twitter*, mas acidental no nível da escolha dos casos. A amostragem probabilística dos casos baseia-se no fato dos eventos de recebimento e transmissão de mensagem de um determinado assunto no *twitter* ser aleatório e imprevisível. Os motivos para a escolha da utilização da amostragem são o tempo necessário e a economia de recursos. Os critérios e o processo realizado estão detalhados a seguir.

Observando-se a dificuldade de prever um efeito viral antes dele acontecer, dentro do período da coleta dos dados, foram intencionalmente identificadas e selecionadas para coleta algumas *hashtags* em uso pela população do estudo. Todos os usuários que, no período da pesquisa, transmitiram as *hashtags* selecionadas tiveram suas informações coletadas. Para definir a amostra foram separadas as *hashtags* que permaneceram no *twitter.*

3.2.2.7 Amostra obtida na pesquisa

A estratégia de obtenção desses dados, bem como outras especificações sobre a forma e conteúdo serão apresentados no capítulo 4. Para fins da descrição da amostra, no entanto, foram antecipadas algumas informações referentes ao procedimento. Neste estudo, portanto, a amostra foi composta pelos usuários do *twitter* que twittaram as *hashtags* quesevaialamafia, lesmakeitawkward e 10pessoasquesonhariaemconhecer no período 10 e 29 de julho de 2013, conforme mostrado na tabela 1. No total foram coletados 849 mil usuários.

Os critérios para seleção dessas *hashtags* foram:

- Contemporaneidade: as *hashtags* deveriam estar sendo difundidas e utilizadas no mesmo momento em que a coleta de dados estivesse operando.
- Viralização: as *hashtags* deveriam ter as características aparentes de um fenômeno viral, tais como propagação espontânea.
- Potencial de Universalidade: as *hashtags* deveriam tratar de fenômenos replicáveis universalmente, de preferencia cultural.

Ao longo da pesquisa, não foi coletada nenhuma *hashtag* diretamente relacionável ao marketing, tais como nomes de marcas, campanhas ou produtos, que atendessem a esses critérios. Mesmo assim, o conceito das *hashtags* coletadas e o comportamento delas pode ser extrapolado para a mercadologia ao passo que todas trataram fenômenos culturais diretamente relacionados aos campos de trabalho do marketing, tais como jogos, personalidades e política.

Entre as *hashtags* desconsideradas na pesquisa, mas que foram previamente analisadas consideraram-se *hashtags* apresentadas em programas de TV e *hashtags* de lançamentos de produtos tais como *podcasts* (que não apresentaram características de viralização) e hashtags que criticavam ou apoiavam grandes produtos ou personalidades (que não se propagaram no período da coleta).

Tabela 1 - Distribuição da amostra segundo *hashtag.*

Hashtag	Breve descrição	Usuários	
		N.	%
#quesevayalamafia	Legenda dada a uma manifestação democrática realizada na Espanha, organizada pelas redes sociais.	321.573	38%
#letsmakeitawkward	Jogo proposto a partir da hashtag, em que usuários do Twitter indicavam 2 pessoas que já haviam se beijado.	160.014	19%
#10pessoasquesonhariaemconhecer	Os membros da rede citavam 10 pessoas interessantes e utilizavam a hashtag para catalogar o grupo.	367.644	43%
Total		849.231	100%

Visto que para reproduzir a rede de um usuário é necessário reproduzir seus vizinhos e partindo da definição que o objeto da pesquisa são os usuários do *twitter* que publicaram as *hashtags* da pesquisa no período de tempo da coleta, tem-se dois tipos de indivíduos coletados: o indivíduo emissor da *hashtag*, ou infectado, e os outros indivíduos que fazem parte da sua rede de vizinhos, mas que não foram infectados pela *hashta*g no período da coleta. Sob este critério, a amostra está distribuída conforme a tabela 2. Na média das três *hastags*, 0,33% dos usuários da amostra foram infectados, sendo que a *hashtag* com maior infecção ficou em 0,39% e a menor com 0,27%.

Tabela 2- Distribuição da amostra entre infectados e vizinhos.

Hashtag	**Não-infectados**		**Infectados**	
	N.	**%**	**N.**	**%**
#quesevayalamafia	320.317	99,61%	1256	0,39%
#letsmakeitawkward	159.477	99,66%	537	0,34%
#10pessoasquesonhariaemconhecer	366.647	99,73%	997	0,27%
Total	846.441	99,67%	2790	0,33%

Já entre os infectados, há aqueles (difusores) cujos seguidores também foram infectados (infectados difusores)e aqueles que apenas divulgaram a mensagem, mas não tiveram desempenho positivo na difusão (infectados não difusores). A tabela 3 mostra essa distribuição. Na média 45,23% dos usuários infectados tiveram sucesso no recontágio de mensagem para pelo menos um usuário, ficando entre 42,52% e 48,79% entre as*hashtags* da amostra.

Tabela 3 - Distribuição da amostra entre infectados difusores e não difusores.

Hashtag	Infectados	Infectados Difusores		Infectados Não Difusores	
	N.	N.	% Total	N.	%
#quesevayalamafia	1.256	534	42,52%	722	57,48%
#letsmakeitawkward	537	262	48,79%	275	51,21%
#10pessoasquesonhariaemconhecer	997	466	46,74%	531	53,26%
Total	2.790	1.262	45,23%	1.528	54,77%

Pela limitação da API do *Twitter* e por fins de agilidade e viabilidade computacional da pesquisa, limitou em 600 o número de vizinhos de cada usuário a serem baixados. Esse número é bem superior ao número médio de seguidores (208) do usuário típico do *Twitter* descrito por Roberts (2012). Essa limitação não interferiu em nada a descrição dos usuários com até 600 seguidores. Por exemplo, se um usuário tinha 208 seguidores, todos os seus seguidores foram baixados para a pesquisa). Se o usuário tinha mais de 600 seguidores, o sistema baixou seus seguidores ordenados dos mais recentes para os mais antigos. Se, mesmo sem ter sido baixado para um usuário específico, um vizinho desse usuário já estava no sistema por conta de downloads de outros usuários, ele era conectado, também a esse usuário, como exemplifica hipoteticamente a tabela 4.

Tabela 4- Operação do limitador de downloads de usuários.

Usuário	Seguidores Reais	Seguidores Baixados anteriormente para os Vizinhos	Seguidores Baixados	Número de Seguidores no banco de dados.	Número de Seguidores no banco de dados.
A	75	7	68	75	75
B	600	32	568	600	600

C	890	32	600	632	632

Dada essa limitação, o número de seguidores real dos usuários não corresponde ao número de seguidores baixados para composição da amostra. Por isso, para calcular as variáveis que envolviam o número de seguidores do usuário, utilizou-se o número de seguidores real, retirado do perfil do usuário no *Twitter*.

3.2.3 DADOS: TIPO E COLETA

O instrumento de coleta de dados foi composto de programas de computador ligados à internet. Foram desenvolvidos dois conjuntos de programas para a pesquisa. O primeiro teve a finalidade de monitorar as emissões das *hashtags* analisadas em toda a rede do *Twitter*, identificando e marcando seus emissores. O segundo teve a função de acessar e coletar as informações dos emissores marcados e das pessoas em suas redes de contato de 1º e 2º níveis. Estas informações foram reunidas pelo sistema em bancos de dados específicos, fora da plataforma *Twitter*.com. A metodologia criada para a coleta, bem como as especificações do instrumento estão detalhados no capítulo 4 desta dissertação.

3.2.4 Computação da *Big Data*: Solução da Exigência de Performance Computacional

A capacidade computacional exigida para construir a rede e calcular as variáveis propostas na quantidade de dados do estudo a partir dos softwares desenvolvidos, no tempo proposto, exigiu a locação de um equipamento de altíssima performance, localizado na cidade de Nova York. Para calcular a Big Data, foram contratadas duas máquinas servidores da empresa DigitalOcean.com, com as especificações no quadro 1.

Características	**Turdus01**	**Turdus03**

Memória RAM	32GB	96GB
HD	320GB	960GB
Sistema Operacional	Ubuntu 12.04 x64bits	Ubuntu 12.04 x64bits
Número de Processadores	12	24

Quadro 1 – Espeficação dos computadores locados para coleta dos dados.

Ao passo que um excelente computador convencional tem 4Gb de memória RAM no ano de 2013, os computadores utilizados extrapolaram essas especificações (96GB e 32GB) para que os cálculos das variáveis fosse possível. Mesmo assim, o tempo total de duração da computação dos indicadores (envolvendo configuração do equipamento e testes iniciais) para todas as variáveis e casos do estudo foi de 14 dias.

3.2.5 Tratamento de dados

Hair et al. (2005) recomenda que os estudos de análise multivariada se iniciem com uma análise minuciosa dos dados coletados, para possibilitar ao pesquisador uma avaliação crítica. Em uma primeira analise superficial dos dados, pôde-se comparar algumas variáveis entre os grupos de usuários infectados e não-infectados, avaliar a variação nas médias, amplitudes, curtoses, assimetrias e desvio padrão de algumas variáveis e identificar a possibilidade de analisar grupos diferentes dentro da mesma amostra. No capítulo 4 são detalhadas algumas das explorações realizadas.

Field (2005) afirma que, para haver analise dos dados, é indispensável que eles sejam preparados para tal. Hair et al (2005) complementa que o início de qualquer técnica multivariada de analise de dados estatísticos é o exame dos dados(em busca da compreensão e ajustes de propriedades estatísticas fundamentais) e das relações entre eles, pois para existir uma interpretação estatística adequada, é de fundamental importância que o pesquisador tenha previamente observado o comportamento desses dados.

Os dados da pesquisa foram examinados seguindo 4 fases distintas recomendadas por Hair et. al (2005): 1) Exame da natureza das variáveis, 2) avaliação e tratamento de dados perdidos na análise, 3) identificação de observações atípicas e 4) avaliação da habilidade dos dados de atender pressupostos estatísticos inerentes às técnicas empregadas.

Os tratamentos recomendados por Fields (2009) para correção de possíveis problemas nos dados são remover os casos, transformar os dados e substituir o valor. Remover o caso só deve ser feito se o pesquisador tiver uma boa razão para concluir que ele não representa a população. Quando a distribuição não for normal, a transformação dos dados deve ser feita. Distribuições assimétricas terão valores atípicos e transformações podem reduzir esse impacto. Não sendo efetiva a transformação, deve ser considerada a possibilidade de substituição, quando o valor estiver distorcendo o modelo e não for representativo da população. Dessas opções, a melhor talvez seja a transformação dos dados, pois reduz o impacto dos valores extremos alterando o valor de todos os dados e não apenas de alguns (FIELDS, 2009).

3.2.6 Analise de Dados

Nesta etapa do estudo, são detalhados os procedimentos de analise aos quais os dados coletados foram submetidos. Para explorar a previsibilidade das variáveis de contágio a partir das variáveis independentes do estudo, na busca pela construção de um modelo significativo, foram utilizadas técnicas de regressão de dados. Tendo em vista que o presente estudo pretende explicar a difusão de informação em redes sociais a partir dos indicadores de características estruturais da rede social, para validar as hipóteses, o modelo do estudo contemplou técnicas estatísticas de dependência que, conforme afirma Hair. Et al (2005, p. 129), “permitem ao observador avaliar o grau de relação entre variáveis dependentes e independentes”. Para que fossem contempladas todas as hipóteses propostas, foram utilizadas especificamente as técnicas de Regressão Logística e Regressão Linear Múltipla, em diferentes etapas da analise, conforme descritos a seguir.

3.2.6.1 Analise de Regressão Logística

Hair et. Al (2005) recomenda a técnica da regressão logística quando são encontradas uma variável não métrica e variáveis dependentes métricas. Inicialmente, o presente estudo pretende analisar a difusão de informações na rede Twitter a um nível binário: difusão ou não. Para tal, utilizada uma variável dependente categórica (não-métrica) indicativa da presença ou ausência do contagio e o conjunto das variáveis independentes métricas indicativas das características da rede.

O processo de decisão para utilização dessa técnica seguiu os estágios sugeridos por Hair et. Al (2005), conforme descritos a seguir:

Estagio 1: Objetivos da Analise Discriminante

Os objetivos de (1) determinar a existência de diferenças estatisticamente significante entre os usuários que foram bem e mal sucedidos no contágio de outros usuários e (2) apontar as variáveis independentes que explicam as diferenças nos dois grupos estão em conformidade com os apontamentos de objetivos que atendem a natureza de abordagem da técnica aos olhos de Hait et. Al (2005).

Estagio 2: Projeto de Pesquisa para Analise Discriminante

De acordo com Hair et. Al (2005, p.219), foram consideradas várias questões para uma aplicação bem-sucedida da analise discriminante, incluindo-se “a seleção da variável dependente e das independentes, o tamanho necessário da amostra para a estimação das funções discriminantes e a divisão da amostra para fins de validação”. A consideração das questões presentes em Hair et. Al (2005) para o presente estudo está descrita a seguir.

- A variável categórica dependente deve atender à especificação de compor grupos excludentes e cobrir todos os casos. A abordagem de extremos polares pose se fazer necessária, caso a variável categórica componha três ou mais grupos.

- A seleção das variáveis independentes foi realizada a partir da pesquisa prévia, que resultou em um modelo teórico a ser testado.
- O tamanho da amostra deve extrapolou o mínimo de 5 observações para cada variável preditora e quantia recomendada, que é acima de 20 observações.
- O menor grupo categórico deve excede o mínimo de 20 observações.
- A amostra também foi dividia aleatoriamente em 2 sub-amostras, uma para desenvolver a função discriminante e a outra para testá-la.

Estágio 3: Suposições da Análise Discriminante

Hair et. Al (2005, p.220), considera que “as suposições-chave para determinar a função discriminante são normalidade multivariada das variáveis independentes e estruturas (matrizes) de dispersão e covariância desconhecidas (mas iguais) para os grupos”, mas pondera que pode-se evidenciar sensibilidade da analise discriminante para essas suposições. O mesmo Hair et. Al, (2005, p. 231) ainda afirma que a regressão logística “não depende dessas suposições rígidas e é bem mais robusta quando tais pressupostos não são satisfeitos”. Mesmo assim, no presente estudo, foram contemplados os seguintes passos e preocupações:

- A classificação pode ser afetada negativamente por matrizes de covariância desiguais. A minimização desse efeito, no estudo, é obtida pelo aumento do tamanho da amostra e pelo uso das matrizes de covariância específicas dos grupos.
- A multicolinearidade das variáveis independentes foi observada, visto que variáveis altamente correlacionadas acrescentam pouco poder explicativo ao modelo.
- As relações não-lineares apresentam reflexo na função discriminante e exigem transformações específicas das variáveis. Também foi realizada a eliminação de observações atípicas para obter a linearidade.

Estágio 4: Estimação do Modelo Discriminante e Avaliação do Ajuste Geral

O método de estimação e o número de funções do modelo foram definidos (Hair et. Al, 2005). “Com as funções estimadas, o ajuste geral do modelo pode ser avaliado de diversas maneiras. Primeiro, escores Z discriminantes, também conhecidos como escores Z, podem ser calculados para cada objeto.” (Hair et. Al, 2005, p.221). O mesmo Hair et. Al (2005) descreve os principais métodos, que foram contemplados no trabalho conforme apresentados nos tópicos a seguir:

- Método Computacional: O autor apresenta dois métodos computacionais passíveis de utilização, de acordo com o interesse do pesquisador: (a) a estimação simultânea, que considera todas as variáveis independentes juntas, computando a função com base em todas as variáveis, desconsiderando o poder discriminatório de cada uma e (b) a estimação *stepwise*, que considera a entrada das variáveis na equação por base em seu poder discriminatório individual. A opção (a) é apropriada quando o pesquisador não tem interesse em resultados intermediários construídos a partir apenas das variáveis mais discriminantes, enquanto a opção (b) é útil quando o pesquisador quer considerar eliminar da função as variáveis menos úteis.
- Significância Estatística: depois de computada a função, é necessário que seja avaliado seu nível de significância. Convencionalmente, a função torna-se significante quando o nível de significância apresenta-se abaixo de 0,05.
- Avaliação do Ajuste Geral: Após a identificação das funções discriminantes, o pesquisador deve atentar-se para a verificação do ajuste geral das funções a partir das seguintes tarefas: (1) calcular os escores Z discriminantes a fim de comparar as observações quanto às variáveis que constituem a função, (2) determinar o tamanho da diferença entre os grupos a partir da comparação das médias dos escores do Z discriminante (centroides) nos grupos diferentes. Diferenças significantes entre os centroides indicam sucesso da analise discriminante.
- Avaliação da Precisão Preditiva de Pertinência ao Grupo: O pesquisador deve tomar medidas para calcular os escores para os quais cada Z discriminante são classificados em cada grupo (os escore de corte), construir as matrizes de classificação dividindo a amostra em um grupo

de análise e um grupo de validação e avaliar os níveis de previsão preditiva em comparação aos dados separados aleatoriamente.

Estágio 5: Interpretação dos Resultados

No caso da existência de significância na função estatística e de uma precisão de classificação aceitável, seguiu-se a sugestão de Hair et. Al (2005) que, para cumprir o processo de interpretação da descoberta e determinar qual a importância individual das variáveis independentes, o pesquisador siga os seguintes métodos:

- Pesos Discriminantes: pesos discriminantes maiores indicam que a variável oferecem maior contribuição para interpretar o fenômeno, e o sinal indica se a contribuição é positiva ou negativa.
- Cargas Discriminantes: Essas cargas refletem o quanto as variáveis independentes estão correlacionadas entre si.
- Valores F parciais: No caso da utilização do método *stepwise*, o pesquisador pode utilizar os valores F para comparar o poder discriminatório das variáveis.
- Rotação das Funções Discriminantes: Após o desenvolvimento das funções, buscando facilitar a interpretação, as funções podem ser rotacionadas.
- Índice de Potência: é um indicativo de poder de cada variável relativamente às outras variáveis do modelo.
- Disposição Gráfica de Cargas Discriminantes: o pesquisador também pode apresentar graficamente as cargas discriminantes de cada variável.

Estágio 6: Validação dos Resultados

Para garantir a validade interna e externa dos resultados, Hair et. Al. (2005) considera essencial que haja a validação cruzada, com dois grupos da mesma amostra ou comparando duas amostras. Outros pesquisadores sugerem que esse procedimento seja repedido várias vezes (Hair et. Al, 2005). Também é recomendado, ainda por Hair et. Al. (2005), após a identificação das variáveis

dependentes que oferecem maior contribuição à discriminação, traçar o perfil de cada um dos grupos, para obter uma melhor avaliação do fenômeno.

3.2.6.2 Analise de Regressão Linear Simples

Em um segundo estágio de análise de regressão, o presente estudo teve, para os casos em que houve difusão, interesse em avaliar relação entre a potência da difusão (difusão relativa) e as variáveis independentes do estudo. Hair et. Al. (2005) recomenda a analise de regressão quando “o objetivo da analise é prever uma única variável dependente a parir de uma ou mais variáveis independentes”.

Seguindo a recomendação de Hair et. Al. (2005), o processo de decisão para a analise de regressão linear simples utilizado foi descrito nas próximas linhas.

Estágio 1: Objetivos da Regressão Múltipla

Como mencionado acima, a analise a seguir está alinhada com Hair et. Al. (2005) em relação ao objetivo. Três questões principais, sugeridas por Hair et. Al. (2005), foram consideradas para verificar este alinhamento, conforme mostrado a seguir nos tópicos adaptados a partir do mesmo autor:

- Adequação do Problema de Pesquisa: Existem duas grandes classes de problemas de pesquisa utilizados nessa técnica, a previsão e explicação. O estudo encaixa-se melhor na classe da explicação, pois pretende, sobretudo “avaliar objetivamente o grau e caráter da relação entre variáveis dependente e independente” (p. 145). Para selecionar as variáveis independentes, o pesquisador deve basear-se na teoria, e pode interpretar a variável estatística baseado na importância das variáveis independentes, nos tipos de relações encontradas ou nas inter-relações existentes entre essas variáveis. No que tange as relações encontradas, na busca de uma relação linear, o autor levanta a possibilidade, por parte do pesquisador, de realizar transformações da variável original.

- Especificação de uma Relação Estatística: o presente estudo encaixa-se na classificação de uma relação estatística, pois as variáveis independentes podem estimar um valor médio da variável dependente, mas não um número exato.
- Seleção das Variáveis Dependente e Independentes: A seleção das variáveis foi realizada a partir dos objetivos e hipóteses da pesquisa, e seguiu as recomendações do autor para evitar erros de medida ou de especificação.

Estágio 2: Planejamento de Pesquisa de uma Análise de Regressão Múltipla

Com o intuito da manutenção dos critérios de significância prática e estatística, o tamanho da amostra e a natureza das variáveis independentes foram avaliadas à luz das orientações em Hair et. Al. (2005). Também foram consideradas, à vista do mesmo autor, a criação de novas variáveis para representar relações entre as variáveis dependente e independente. As considerações tomadas no estudo, sumarizadas de Hair et. Al. (2005), estão apresentadas nos tópicos abaixo:

- Poder Estatístico e tamanho da amostra: Visto que “o tamanho da amostra tem um impacto direto sobre a adequação e o poder estatístico da regressão” (p. 147), refletindo em sensibilidade excessiva em amostras muito grandes e necessidade de relações muito fortes para amostras muito pequenas, para o estudo proposto (envolvendo 6 variáveis independentes, esperando-se um nível de significância de 0,05, com uma amostra acima de 500 casos) pretenderemos detectar um R2 mínimo de 5%.
- Generalização e Tamanho da Amostra: O autor apresenta uma regra geral, em que a razão entre observações e variáveis independentes não deve nunca exceder a proporção de 5 para 1, enquanto os níveis desejados estão entre 15 e 20 para 1. No presente estudo, buscou-se extrapolar estas premissas a um nível acima de 50 para 1.
- Previsores de Efeitos Fixos Versus Aleatórios:
- Criação de Variáveis Adicionais: Hair et. Al (2005) reconhece que “a falta de habilidade da regressão de modelar relações não-lineares diretamente

pode restringir o pesquisador quando ele enfrenta situações nas quais uma relação não-linear é sugerida pela teoria ou detectada quando examinamos os dados" (p. 149). Para avaliar isso, foram considerados procedimentos de transformação das variáveis conforme sugeridos pelo autor e especificados a seguir: (a) representação de efeitos curvilíneos com polinômios: onde criamos novas variáveis para modelar componentes curvilíneos da relação com a variável dependente a partir de transformações de potenciais de uma variável independente. (b) representação de efeitos moderadores: foram levados em consideração as possibilidades de uma variável independente exercer efeito de moderação em outra variável independente. Estes procedimentos, conforme sugerido por Hair et. Al (2005), tiveram os desempenhos de seus efeitos avaliados pelo pesquisador em uma comparação antes *versus* depois, e a devida atenção foi dada ao risco de perda de generalização do estudo com a aplicação desses métodos.

Estágio 3: Suposições em Analise de Regressão Múltipla

Considerando as transformações acima e as características das variáveis, Hair et. Al (2005) sugere várias suposições a fim de garantir a inexistência de relações que afetam o procedimento estatístico. As suposições consideradas no estudo são detalhadas abaixo, sumariadas a partir da produção em Hair et. Al (2005).

- Linearidade do fenômeno: A linearidade da relação entre as variáveis está associada ao grau a que a variação das variáveis dependentes e independentes está associado. Neste estudo, a linearidade foi avaliada através de gráficos de resíduos.
- Variância constante do termo de erro: A presença de variâncias desiguais foi avaliada a partir do teste Levene.
- Independência dos termos de erro: para garantir que o valor previsto não estará relacionado com qualquer outra previsão, foi avaliado o gráfico de resíduos, em busca de um padrão nulo.

- Normalidade da distribuição dos termos de erro: A normalidade foi buscada a partir de testes de normalidade e pela avaliação do histograma, em busca de uma distribuição próxima da normal.

Estágio 4: Estimação do Modelo de Regressão e Avaliação do Ajuste Geral do Modelo

Após os estágios de especificação de objetivos da analise, seleção das variáveis e verificação das suposições, o modelo em si foi avaliado relação à precisão preditiva das variáveis, buscando uma máxima previsão, conforme sugerido por Hair et. Al (2005) e apresentado sumariamente a seguir.

- Tratamentos Gerais para Seleção de Variáveis: Para selecionar as variáveis do modelo entre as variáveis apresentadas na teoria e as sugeridas pelo pesquisador, foram experimentados e combinados os tratamentos a seguir: (a) especificação confirmatória: o pesquisador tem autonomia para especificar as variáveis componentes do modelo, a partir da teoria; (b) métodos de busca sequencial: em que foram seletivamente adicionadas e retiradas variáveis em busca das medidas de critério propostas (foram testados os métodos *forward* e de adição *forward* e eliminação *backward*);
- Exame de Significância Estatística do Modelo: para verificar se o modelo preditivo pode realmente representar a amostra, os seguintes passos foram seguidos: (a) significância do modelo geral: avaliou-se se o R2 é maior que zero para avaliar se o modelo proposto opera melhor que uma avaliação aleatória sem as variáveis; (b) testes de significância dos coeficientes de regressão: visto que a analise é baseada em uma amostra e não em um senso, o coeficiente de regressão foi validado pelo tamanho da amostra.
- Identificação de Observações Influentes: foram identificadas e avaliadas quanto ao impacto as variáveis do modelo que poderiam influenciar fortemente o modelo de regressão, a partir da analise de resíduo. As observações verdadeiramente excepcionais foram eliminadas da analise.

Estágio 5: Interpretação da Variável Estatística do Modelo

De acordo com Hair et. Al (2005, p. 164), “o pesquisador deve avaliar não apenas o modelo de regressão estimado, mas também as variáveis independentes potenciais que foram omitidas se buscas sequenciais ou analises combinatórias foram empregadas”. A fim de avaliar as significância geral e estatística do modelo, foram avaliados os coeficientes estimados e o impacto potencial de variáveis omitidas, de acordo com as orientações de Hair et. Al (2005), sumariadas a seguir:

- Utilização e Padronização dos Coeficientes de Regressão: Para eliminar o problema de lidar com diferentes unidades de medida, os coeficientes beta foram originados a partir dos dados padronizados.
- Avaliação da Multicolinearidade: para avaliar o grau a que as variáveis independentes estão correlacionadas entre si (multicolinearidade) e determinar seu impacto sobre o resultado, bem como as ações corretivas necessárias, foram abordados no trabalho: (a) os efeitos da multicolinearidade: na explicação da regressão, na estimação dos coeficientes de regressão e nos testes de significância estatística; (b) a identificação da multicolinearidade: foi avaliado no estudo o cálculo da correlação direta, já que as variáveis estatísticas são elaboradas a partir de apenas uma variável independente (foram considerados índices altos a partir de 0,9); (c) ações corretivas para a multicolinearidade: considerou-se no estudo omitir variáveis independentes para eliminar a multicolinearidade, se necessário.

Estágio 6: Validação dos Resultados

Para garantir os potenciais de generalização e transferibilidade do modelo de regressão, seguimos a especificação de Hair et Al (2005), que sugere que a amostra seja particionada para testar a representatividade do modelo ou então que o modelo seja testado em uma outra amostra.

4 METODOLOGIA DE COLETA DE DADOS

Para efetivação da pesquisa, utilizou-se uma estratégia pela qual os dados secundários coletados do banco de dados do *Twitter* foram acessados, transferidos e organizados em uma base externa ao *Twitter*.com. Essa coleta, por sua vez, demandou grande capacidade computacional e um banco que permitisse analise relacional. A seguir é aprofundada a descrição desses dois pontos.

Nascida na década de 80, a tecnologia da base de dados de Grafos (Graph Database) ganhou força com o surgimento de grandes redes, pela necessidade de gerenciar informações cujos esquemas e instâncias são naturalmente modeladas como os grafos da matemática ou generalizações deles (ANGLES & GUTIERREZ, 2005).

Seguindo a visão de Codd (1980), em que um modelo de base de dados deve atender aos critérios conceituais (de estruturação, manutenção e descrição de dados) e práticos (de suportar as metodologias de design da base de dados), para essa análise, utilizaremos uma *Graph Database*, pois tão importante quanto os dados em si, no caso do trabalho presente, é a relação entre esses dados. Sob esse critério, esse tipo de base de dados tem grande validade, pois de acordo com Kachiola et al. (2005), ela pode ser modelada como um grafo direto, em que os nós são entidades e os laços entre eles são relacionamentos ou características, provendo assim o pesquisador de uma maneira geral de modelar uma variedade de relações entre os dados em si de forma mais rápida, comparada com outros tipos de bancos de dados (Huan et. Al, 2004). Para essa pesquisa, adotou-se o modelo de banco de dados Neo4j versão 1.8.2, de fevereiro de 2003, fornecida pela *Neo Technology*, que possuí uma edição de licença comum, identificada como GPLv3, e opera na linguagem Java. Essa opção de banco já possuí nativamente um navegador visual, um mecanismo de processamento, através da linguagem de pesquisa *Cypher*, e um mecanismo de armazenamento dos dados (Neo4j, 2013).

Para Angeles & Gutierrez (2005), os fundamentos dos modelos de base de dados Graph estão na representação de entidades (algo que existe no universo como uma unidade simples e completa) e relações (propriedade que estabelece a conexão entre duas ou mais entidades). Para cumprir a proposta da pesquisa, o banco é composto das seguintes entidades e relações:

Entidades	Relações
Usuários Análises Referências Tweets	Conexões

Quadro 2– Composição do banco de dados da pesquisa: entidades e relações.

Além das relações citadas acima, as entidades componentes do banco também recebem os seguintes atributos:

Entidade	Atributos
Usuário	Nome Hashtag que Twittou Número da conta Twitter Número de seguidores. Número de Seguidos
Analysis	Identificação da análise na base de dados
Reference	Alguma menção possível na base.
Tweet	Conteúdo da mensagem enviada.

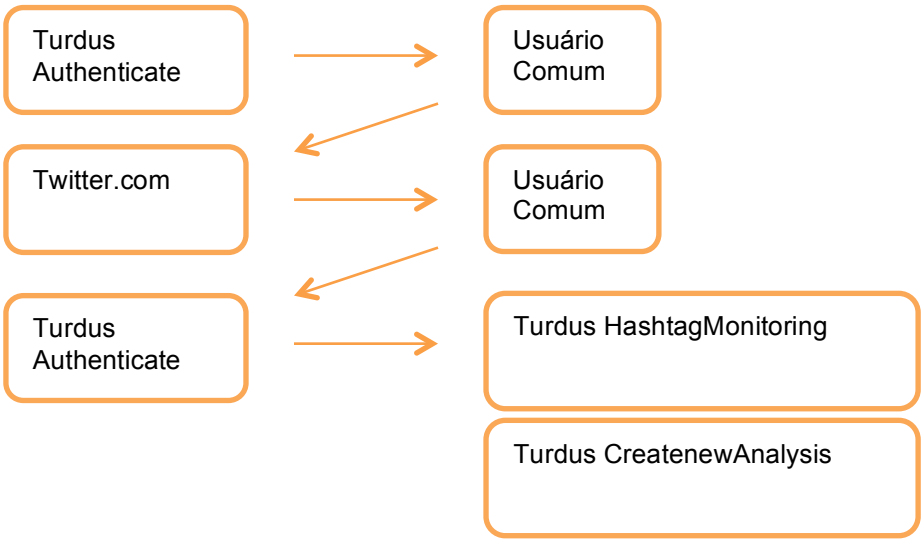
Quadro 3- Composição do banco de dados da pesquisa: atributos das entidades.

Dados o grande volume de dados almejados e as limitações oferecidas pela API do *Twitter*, foi necessário o desenvolvimento de um conjunto de programas de computador específicos, que respeitassem as regras impostas pela plataforma e acessassem as informações necessárias, simulando o comportamento de um usuário comum. Esse conjunto foi batizado de *Turdus* e está descrito em três partes: *Turdus Authenticate, Turdus HashtagMonitoring* e *Turdus CreateNewAnalysis.*

4.1 *TURDUS AUTHENTICATE*: SISTEMA DE AUTENTICAÇÃO DE USUÁRIOS

O ***Turdus Authenticate***, primeiro programa utilizado na estratégia de obtenção de dados, não tinha por fim executar qualquer tipo de acesso ao banco de dados do Twitter, mas sim autenticar os programas *HashTagMonitoring* e

CreateNewAnalysis com as identidades de acesso de voluntários usuários comuns da plataforma, atribuindo aos programas identidades com logins e senhas de pessoas comuns, para que eles pudessem, de acordo com as regras e limitações do Twitter descritas anteriormente, ter suas solicitações de informações atendidas, sem comprometer as informações de *login* e senha dos usuários a um sistema externo ao Twitter.com. A rotina de funcionamento de programa está ilustrada na figura 7.



1) Solicitação de Autenticação: Ao ser acionado pelo usuário, o *Turdus Authenticate* abre a janela de autenticação do *Twitter*.com (com campos para login e senha) no navegador do usuário.
2) *Login* e Senha: O usuário informa seu *login* e senha ao *Twitter*.com, acessando sua conta.
3) Obtenção de Token: O *Twitter*.com informa ao usuário um número de *token* de acesso.
4) Entrega de Token: O usuário voluntariamente entrega ao Turdus Authenticate o número de *token* de acesso.
5) Atribuição de Token: O Turdus Authenticate atribuí aos programas Turdus HashtagMonitoring e Turdus CreateNewAnalysis o *token* do usuário voluntário.

Figura 6 - Rotina de funcionamento do programa *Turdus Authenticate.*

A rotina mais detalhada do algoritmo do *Turdus HashtagMonitoring*está descrita em forma de algoritmo na figura abaixo:

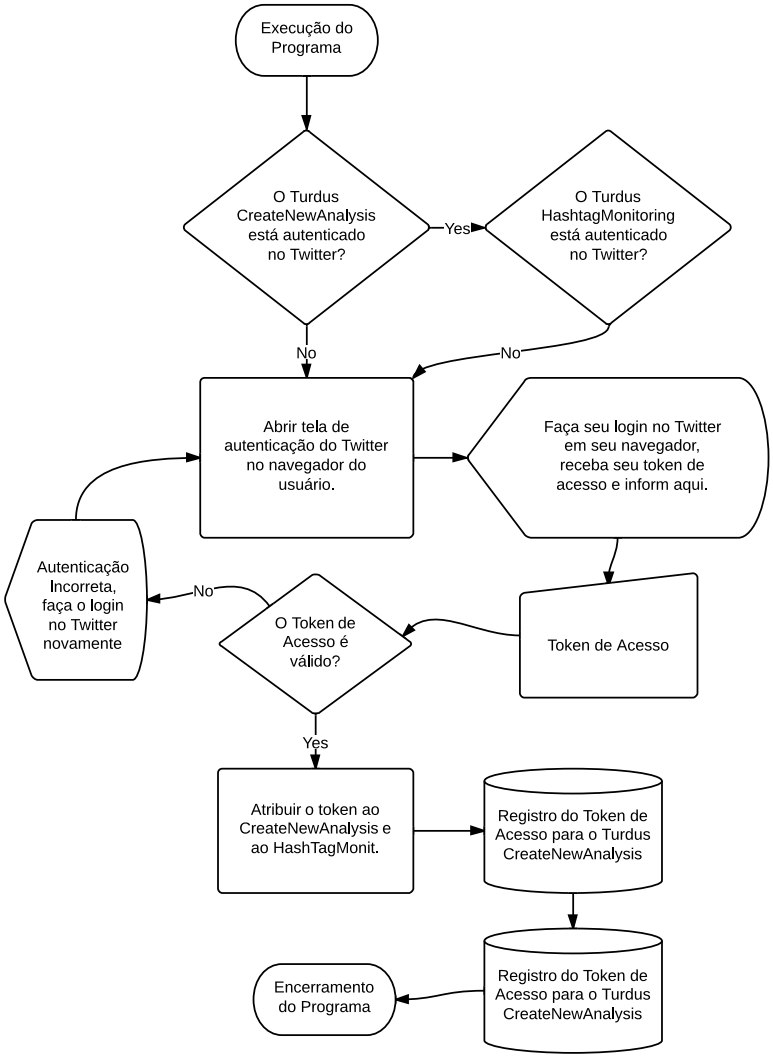
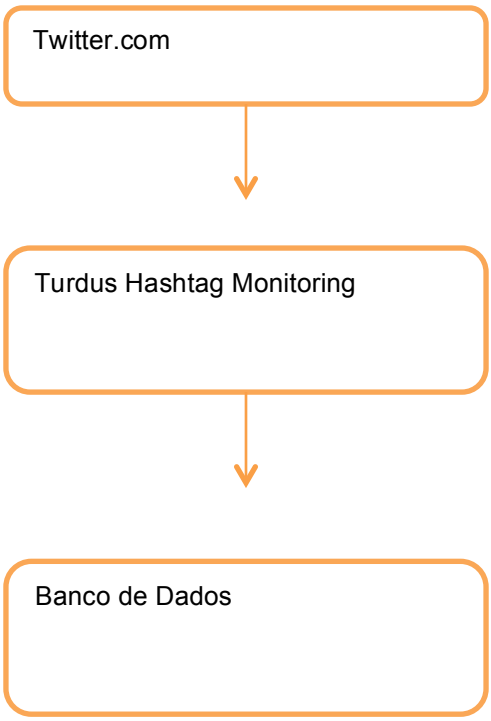


Figura 7 – Algoritmo detalhado do programa T*urdus Authenticate.*

4.2 TURDUS HASHTAGMONITORING: SISTEMA DE MONITORAMENTO DAS *HASHTAGS* SELECIONADAS

O ***Turdus HashtagMonitoring***, segundo programa utilizado na estratégia, identifica as mensagens enviadas no *Twitter*.com relevantes à pesquisa e incluí no banco de dados o conteúdo da mensagem, a identificação do usuário e o momento (*timestamp*) em que a mensagem foi enviada. Entre os milhões de mensagens públicas enviadas pelo *Twitter* diariamente, o interesse dessa pesquisa restringe-se às mensagens enviadas com os *hashtags* selecionadas. Para separar essas

mensagens e adicionar os casos emissores ao banco de dados da pesquisa, a rotina de funcionamento do programa está descrita na rotina a seguir.



1) Recebimento de mensagens: O sistema Turdus Hashtag Monitoring recebe todas as mensagens abertas que são enviadas no *Twitter*.com no mundo.
2) Processamento de mensagens e filtragem de usuários: O sistema busca no conteúdo das mensagens as *hashtags* estabelecidas para a pesquisa.
3) Marcação e Inserção dos usuários na Base de Dados: O sistema insere os usuários filtrados na etapa 2 na base de dados da pesquisa e marca cada um esses usuários com o termo “infected”.

Figura 8 - Rotina de funcionamento do programa *Turdus Hashtag Monitoring.*

A rotina mais detalhada do algoritmo do Turdus HashtagMonitoring está descrita em forma de algoritmo na próxima figura:

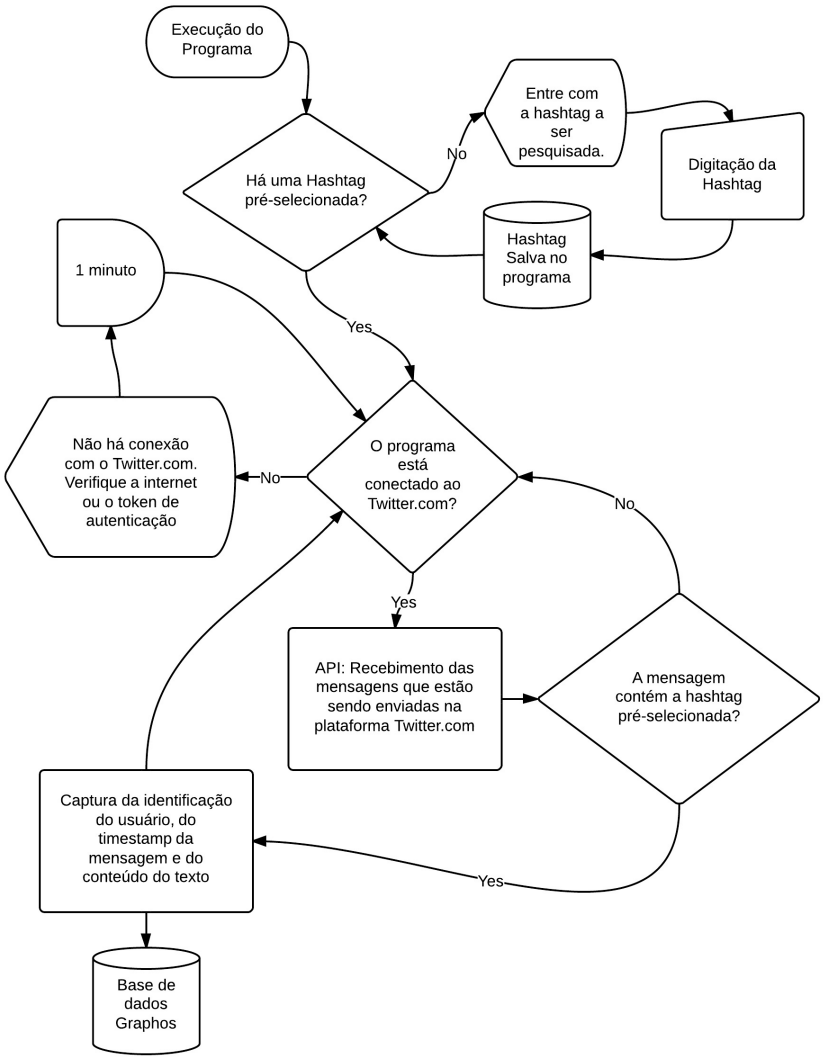
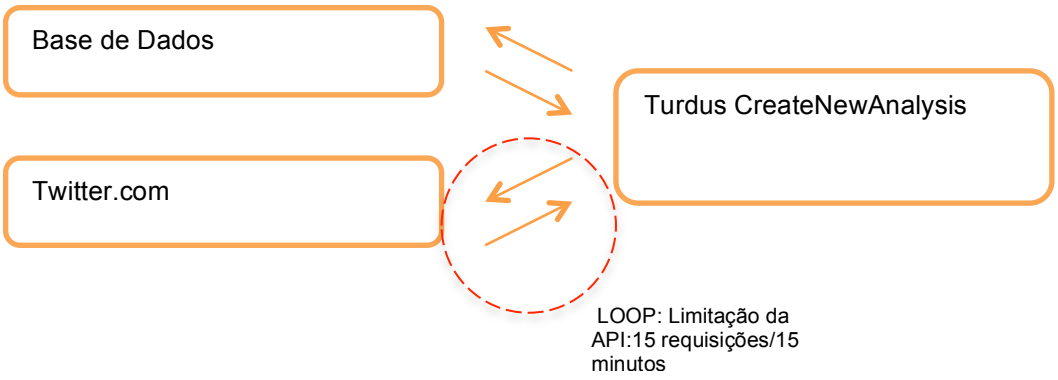


Figura 9 – algoritmo detalhado do programa *Turdus Hashtag Monitoring.*

4.3 TURDUS CREATENEWANALYSIS: SISTEMA DE CONSTRUÇÃO DE BASE DE DADOS DA PESQUISA

O ***Turdus CreateNewAnalysis***, terceiro programa utilizado na estratégia, constrói a base de dados da pesquisa, detalhando os casos inseridos no banco pelo *Turdus HashtagMonitoring* com as informações necessárias para as análises, buscando-as na base de dados do *Twitter*.com, de acordo com o ritmo permitido pela *Twitter* API (15 requisições a cada 15 minutos). Para obter as informações necessárias para complementar a base de dados, a rotina de funcionamento do programa está descrita na rotina a seguir.



1) Seleção dos Alvos: O sistema procura os casos inseridos na base pelo *Turdus HashtagMonitoring*.
2) Solicitação de Informações: O sistema solicita à API do *Twitter*.com as informações referentes a esses usuários definidas na pesquisa.
3) Recebimento de Informações: A API do *Twitter* entrega ao sistema as informações solicitadas na etapa anterior.
4) Alimentação do Banco: O sistema complementa a base de dados da pesquisa com as informações necessárias.
5) Restrição de Passo: Para respeitar a limitação exigida pela *Twitter* API, o processo de requisições inicia-se e pausa, de 15 em 15 requisições, a cada 15 minutos.

Figura 10 - rotina de funcionamento do programa *Turdus CreatenewanAlysis.*

A rotina mais detalhada do algoritmo do *Turdus Hashtag Monitoring* está descrita em forma de algoritmo na figura abaixo:

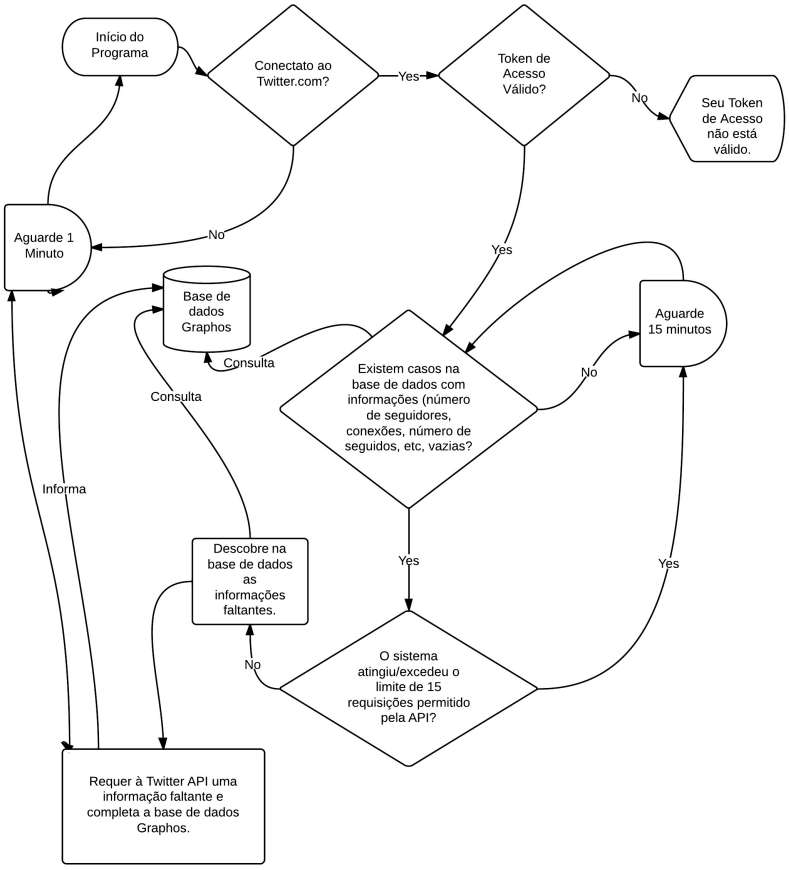


Figura 11 – algoritmo detalhado do programa Turdus CreatenewanAlysis

Depois de construído o banco de dados para a pesquisa, foram criadas, a partir da disposição dos conjuntos de nós e laços do banco de dados, as variáveis da pesquisa. Para isso, desenvolveu-se outro programa chamado **Turdus Compute Indicators**, que percorre o banco de dados grafos calculando as variáveis do estudo e exportando-as para um arquivo de valores separados por vírgula (CSV).

Título da Coluna	Descrição
Index	Número índice do caso na base de dados.
screen_name	nome do usuário no twitter
isInfected	condição de o usuário twittou a hashtag ou não (TRUE/FALSE)
Seguidores	N de seguidores do usuário.
Seguidos	N de seguidos do usuário.
Popularidade_Relativa	variável Popularidade_Relativa
Degree	variável Degree
Clustering_N	variável Clustering_N
Contagio	variável Contagio
Contagio_Relativo	variável Contagio_Relativo
Clustering_M	variável Clustering_M
Eigenvector_Centrality	variável Eigenvector_Centrality
Hashtag	hashtag twittada pelo usuário.

Quadro 4 - Lista de variáveis geradas pelo *Turdus Compute Indicators* e os algoritmos gerados.

5 APRESENTAÇÃO E ANALISE DOS RESULTADOS

Este capítulo apresenta em maiores detalhes as características da população do estudo e, a partir da aplicação da metodologia apresentada nos capítulos anteriores, apresenta e discute os resultados obtidos. Em um primeiro momento, a base de dados passou por uma etapa de análise e tratamento. Depois, a análise estatística.

5.1 ANALISE DESCRITIVA DA POPULAÇÃO

Por ser uma plataforma aberta, vários levantamentos da plataforma Twitter.com foram realizados por diversos pesquisadores e empresas de pesquisa. A partir deles, pode-se obter um maior aprofundamento da população da plataforma.

5.1.1 O *Twitter:* formato e operação.

O *Twitter* tornou-se uma das redes de maior sucesso em todo o mundo. Lançado em 21 de março de 2006, o site*twitter.com* é uma plataforma de rede social que permite que usuários enviem mensagens de até 140 caracteres, de conteúdo livre, conhecidos como “tweets”. Essas mensagens podem ser enviadas, por meio de aplicativos (de celular, tablets e outros) ou pela interface de usuário no site, a qualquer pessoa interessada (através do acesso à página do perfil do usuário emissor) ou a qualquer pessoa que tenha optado por seguir (o seguidor) o perfil do usuário emissor. O conteúdo das mensagens é muito variado e a ferramenta oferece ao usuário uma possibilidade de catalogação do conteúdo da mensagem através da utilização de *hashtags*. Além de “tweetar” (termo que denomina a ação de enviar tweets para a rede), os usuários também podem atualizar informações de seu perfil, como sexo, idade e breve descrição, e enviar mensagens diretas privadas a outros usuários.

5.1.2 Os Usuários do Twitter

ROBERTS (2012) Afirma que o usuário típico do *Twitter* é uma mulher jovem, em um iPhone, com 208 seguidores. Dos 500 milhões de usuários cadastrados na plataforma, de acordo com Schroeder (2003), 200 milhões estão ativos mensalmente. BEEVOLVE (2012) considera que 53% dos usuários da plataforma são mulheres, e que enquanto um homem envia 567 tweets na rede, uma mulher envia 610.

É difícil estimar a distribuição etária dos usuários na plataforma, já que apenas 0,45% dos usuários divulga a própria idade na rede. Dos usuários que divulgaram a informação, 73,7% têm entre 15-25 anos (BEEVOLVE, 2012). Mas o mesmo BEEVOLVE (2012) ressalva que essa informação não deve corresponder à real distribuição etária da plataforma, já que os usuários entre 15-25 anos são os mais interessados em divulgar a própria idade na rede. Ainda segundo Beevolve (2012), apenas 30% dos usuários informou algum tipo de "bio" (descrição pessoal) em seu perfil.

Quanto ao número de "seguidores" e "seguidos", os usuários do *twitter* apresentam uma grande variação. Mesmo o usuário médio tendo 208 seguidores, podemos encontrar usuários com centenas de milhares de seguidores e seguidos. Os gráficos 1 e 2 mostram a distribuição do número seguidores e seguidos do *twitter* por faixa de número de seguidores. 74,1% e 81,1% dos usuários tem até 50 seguidos e seguidores, respectivamente (Beevolve, 2012).

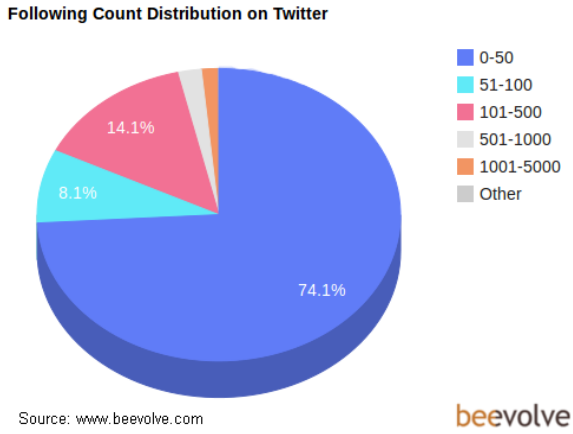


Gráfico 1 - distribuição do número de seguidores do Twiter.
Fonte: Beevolve, 2012,

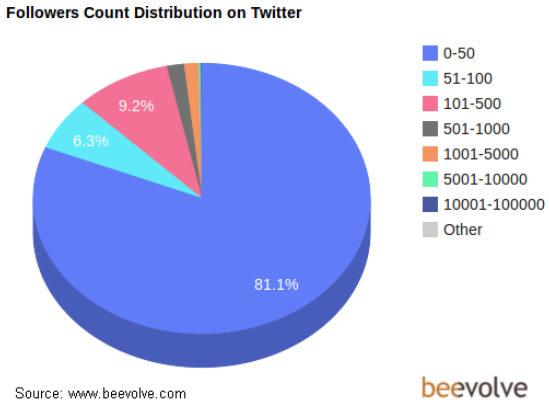


Gráfico 2- Distribuição do número de seguidos do Twiter.
Fonte: Beevolve, 2012.

Entre os países com maior quantidade de usuários no *Twitter*, o Brasil ocupava em 2012 a segunda posição com mais de 40 milhões de usuários ativos e não ativos, de acordo com o rank fornecido pela GlobalWebindex, antecedido por Estados Unidos, mais de 140 milhões e precedido pelo Japão com quase 35 milhões.

Chama atenção nesse rank a ausência da China, contrária a informação de Lipman (2013) de que possui o maior número de usuários ativos. Segundo ele, os usuários são forçados desenvolver meios de burlar a grande muralha cibernética para poder acessar a plataforma. Os países com maior número de usuários ativos, segundo o mesmo Lipman (2013) (gráfico 4) são, na ordem, China (35,5 milhões), Índia (33 milhões), Estados Unidos (22,9 milhões) e Brasil (19,6 milhões).

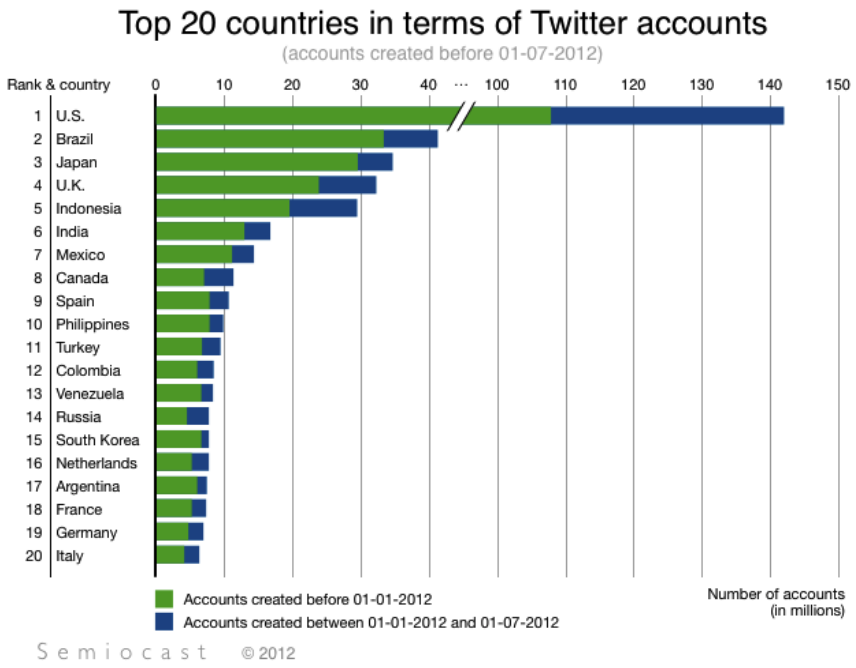


Gráfico 3- Países com maior número de contas no Twitter.
Fonte: Semiocast, 2012.

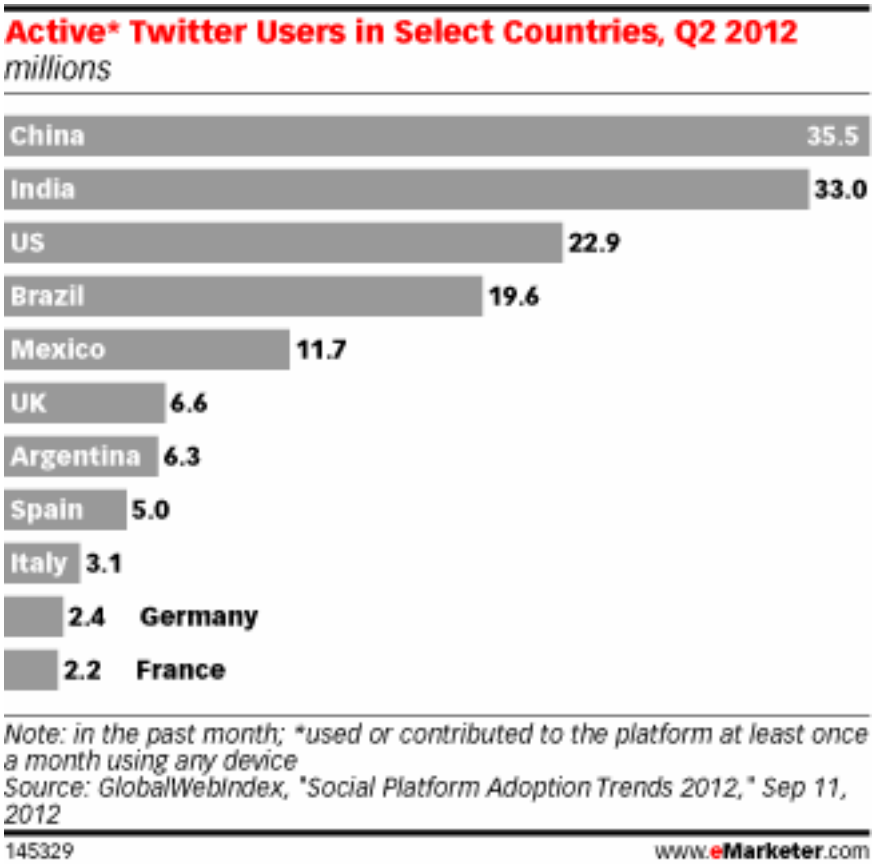


Gráfico 4- Países com maior número de usuários ativos no Twitter.
Fonte: marketer.com, 2012.

Beevolve (2012) aponta uma variação de distribuição de gênero na plataforma: se comparados países desenvolvidos e em desenvolvimento, os países desenvolvidos têm uma distribuição mais igualitária entre os gêneros, com maior número de mulheres entre os usuários, enquanto os países em desenvolvimento agregam mais homens à plataforma (Gráfico 5).

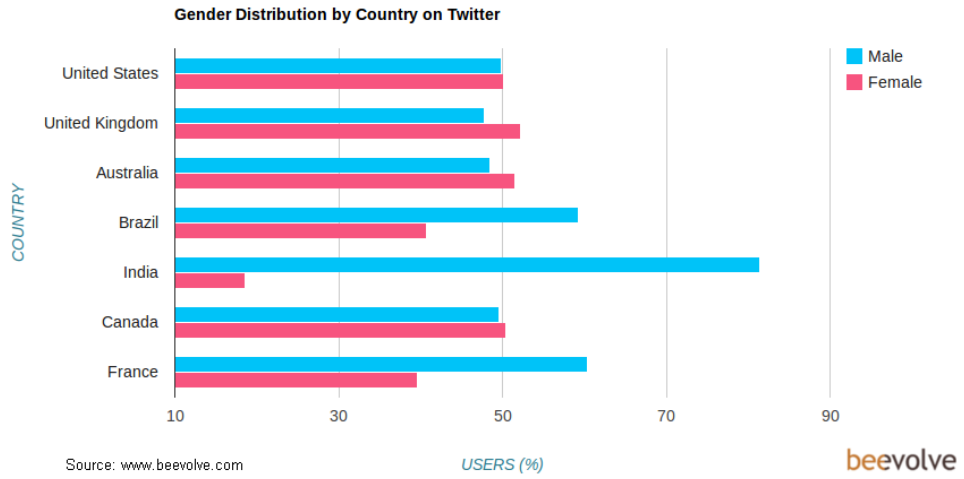


Gráfico 5– Distribuição do gênero por país no Twitter,
Fonte: Beevolve, 2012.

5.1.3 Os Tweets

Os Tweets são as mensagens de até 140 caracteres que todo usuário do *Twitter* pode enviar para seus seguidores ou para as pessoas que visitam o seu perfil. Estas mensagens, desde o lançamento da plataforma em 2006, cresceram em volume e contabilizavam em março/2012, segundo *TWITTER* BLOG (2012), 340 milhões por dia, provenientes de 140 milhões de usuários ativos. Schroeder (2013) observa que, já em 2013, os tweets chegaram à quantidade de 500 milhões por dia, vindos de 200 milhões de usuários.

Os temas das mensagens são variados e abrangem Tecnologia, Família, Educação, Empreendedorismo, esportes entre outros. O gráfico 5 apresenta as 10 categorias mais abordadas nas mensagem do *twitter*, de acordo com Beevolve (2012).

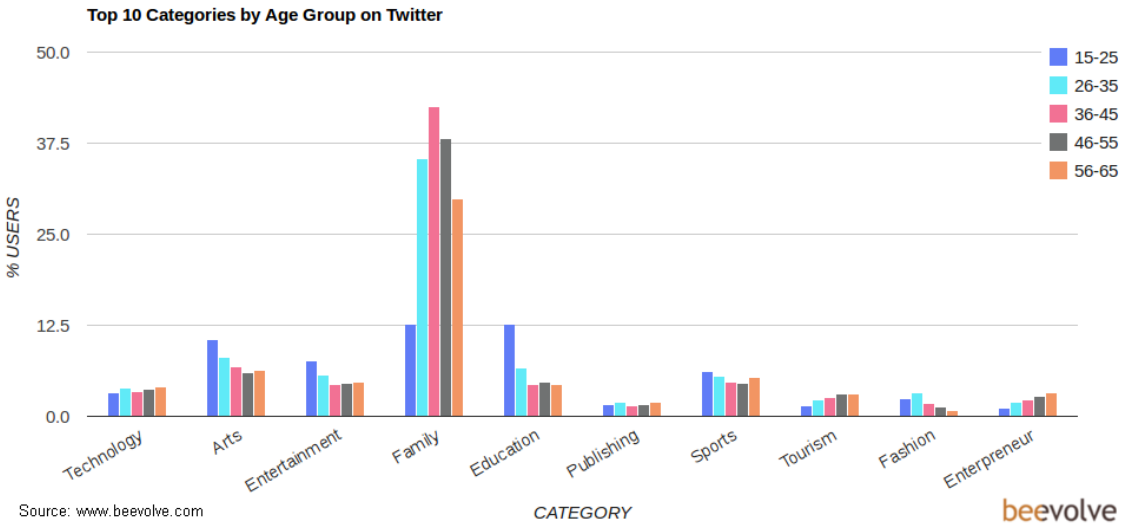


Figura 12 – Categorias mais discutidas no Twitter por faixa etária.
Fonte: Beevolve, 2012.

Por serem de caráter variado, a organização dessas mensagens se dá pela utilização de marcações ou etiquetas (*tags*) de catalogação, conhecidas como *hashtags*. Os próprios usuários que enviam a mensagem podem incluir as *hashtags* com as quais gostariam de catalogá-las. A partir dessas *tags*, outros usuários podem – por mecanismos de busca do *twitter* - encontrar os assuntos que desejam acompanhar.

5.1.4 As *Hashtags*

Por serem desenvolvidas e gerenciadas pelos próprios usuários, as *hashtags* não seguem uma lógica de catalogação, e muitas vezes definem os assuntos tratados no momento através de piadas ou de siglas e abreviações utilizadas pelas pessoas familiarizadas com o assunto. Algumas *hashtags* representam campanhas publicitárias, outras podem representar campanhas políticas, jogos de comportamento no *twitter* ou então designar algum produto cultural (quadro 5).

Rank	Trend	Tempo Total no Top10	Explicação	Categoria da Discussão
1	#HabitsThatAreHardToBreak	1d4h40m	Pessoas discutindo hábitos que são difíceis de mudar.	Comportamento
2	#KanyeWestIsTheKindOfGuyThatWill	10h40m	Críticas ao ego inflado do astro POP Kanye West	Produtos Culturais
3	#cevapver	11h0m	Menções aos protestos contra as medidas autoritárias do primeiro-ministro na Turquia.	Política
4	#1YearOfBelieve	11h5m	Menções à comemoração de 1 ano da publicação do disco acústico "Believe", de Justin Bieber.	Produtos Culturais
5	#*Twitter*osConLosQueQuieroUnaFotoAbrazo	12h45m	Pessoas, em espanhol, discutindo os colegas de *Twitter* com quem gostariam de tirar fotos abraçados.	Comportamento
6	#BugünMilyonlarKAZLIÇEŞMEDE	11h25m	Menção e convites a uma grande manifestação em prol da democracia na Tirquia.	Política
7	Happy Father's Day	20h35m	Comemoração do dia dos pais, seguido de mensagens e homenagens.	Datas Comemorativas
8	#BilDiyeSöylüyorum	8h55m	Não Identificado	Não Identificado
9	#conlosanosaprendi	7h40m	Pessoas, em espanhol, discutindo o que aprenderam com o passar dos anos.	Comportamento
10	#BugünGünlerdenANKARA	8h5m	Menção e convites a uma grande manifestação em prol da democracia na Turquia.	Política
11	#YeniBirDünya	7h5m	Não Identificado	Não Identificado
12	#SoyUnCompradorCompulsivoDe	6h55m	Pessoas, em espanhol, discutindo produtos que compram compulsivamente.	Comportamento
13	#ChupaDilma	6h55m	A presidente do Brasil, Dilma Roussef foi vaiada na abertura da copa das confederações.	Política
14	#*Twitter*DelileriTakipleşiyor	6h5m	Menção e convites a uma grande manifestação em prol da democracia na Tirquia.	Política

Continua

Quadro 5 – *Hashtags* mais discutidas no mundo entre 14/jun e 14/jul/2013.
Fonte: Whatthetrend, 2013).

Continuação

Rank	Trend	Tempo Total no Top10	Explicação	Categoria da Discussão
15	#Is1DLarryRealOrFake	5h00m	Discussão sobre a real situação conjugal de um dos músicos da banda OneDirection.	Produtos Culturais
16	#TürkiyeSokakta	5h10m	Menção e convites a uma grande manifestação em prol da democracia na Tirquia.	Política
17	#SiYoFueraPolicía	5h45m	Pessoas, em espanhol, discutindo quais seriam suas atitudes se fossem policiais.	Política
18	#UnRecuerdoDeLosSimpsons	5h55m	Pessoas lembrando situações da série Os Simpsons.	Produtos Culturais
19	#PolisEvineDön	5h05m	Não Identificado	Não Identificado
20	#invalsi	3h55m	Italianos discutindo a educação secundária no país.	Educação

Quadro 5 – *Hashtags* mais discutidas no mundo entre 14/jun e 14/jul/2013. Fonte: Whatthetrend, 2013).

5.2 ANÁLISE DESCRITIVA DA AMOSTRA

A tabela 5 demonstra que todas as variáveis dessa amostra apresentam dispersão considerável em relação à média. As médias e medianas encontram-se em pontos diferentes. A *trimmed mean*, que é o valor resultado da média ao serem retirados os 5% casos extremos em ambos os lados da distribuição. A comparação da *trimmed mean* com a média indicam que os casos extremos comprometem os resultados observados (PALLANT, 2001).

Os valores de assimetria estão acima de zero para todas as variáveis. Como a média é sempre maior que a mediana, a assimetria de todas as variáveis encontra-se para a direita. Os valores das curtoses de todas as variáveis são maiores que zero, indicando que as distribuições de todas as variáveis têm caudas mais longas que a de uma distribuição normal.

Tabela 5– Estatísticas descritivas da amostra.

	Seguidores	Seguidos	Popul. Relativa	Degree	Cluste-ring_N	Cluste-ring_M	Eigenvector Centrality	Contagio
Média	965,2	663,6	3,661	3,031	0,0660	0,0625	0,001740	1,90
Limite inferior*	861,7	611,2	2,403	1,694	0,0635	0,0553	0,001625	1,69
Limite superior*	1068,8	716,0	4,919	4,369	0,0685	0,0697	0,001856	2,11
5% *Trimmed Mean*	542,1	489,6	1,048	0,913	0,0580	0,0240	0,001269	1,10
Mediana	295,0	311,0	0,930	0,805	0,0428	0,0000	0,000537	,00
Desvio Padrão	2787,6	1410,1	33,9	36,0	0,0678	0,1937	0,003107	5,66
Mínimo	1	2	0,0476	0,0040	0,0000	0,0000	0,000000	0
Máximo	54888	36264	1047,5	1604,9	0,9000	1,0000	0,047397	133
Assimetria	9,3	12,3	21,5	34,9	2,7	3,7	4,7	10,7
Curtose	121,5	237,5	547,7	1451,0	15,0	13,7	38,3	174,3

*** Intervalo de confiança (95%)**

Não existiram dados faltantes (*missing values*) no levantamento de dados.A análise de respostas fora de padrão ou observações atípicas foi realizada inicialmente com a utilização de análise univariada mediante o auxílio do gráfico *boxplot* do SPSS. Posteriormente, foi feita a análise bivariada e multivariada de *outliers*, com a análise de regressão, após definição do modelo de mensuração.

5.3 ANÁLISE DE *MISSILING VALUES* E *OUTLIERS* E TRATAMENTO DOS DADOS

Como os dados foram obtidos através de programas, ocorreram apenas 4 casos de dados faltantes, em que no momento do download das características dos usuários do *twiter*, os casos estavam com nenhum seguidores, ficando a amostra com 2.785 casos.

A análise de *outliers* na dimensão univarida foi realizada, além das estatísticas, de assimetria e curtose, pelo exame dos gráficos de histograma das variáveis e gráfico boxplot. Esse exame identificou muitos casos como outliers, especialmente os com maior número de seguidores. Mas, baseado nas observações de Fieds (2009), não há justificativa para considerá-los não pertencentes a população ou erros na obtenção dos dados, visto que os valores foram obtidos por sistema automatizado e são característico da amostra.

Para tratamento dos dados obtidos, com a ausência de justificativas para substituição ou remoção de *outliers*, fez-se a opção pela alteração da elasticidade das variáveis, em consonância com as recomendações de Field (2009), transformando os dados com a aplicação do logaritmo e da raiz quadrada.

A tabela 6 mostra os resultados do desvio padrão, assimetria e curtose após esse tratamento de dados. Observa-se que a transformação diminui o desvio padrão e melhorou os valores de assimetria e curtose para todas as variáveis. Para as variáveis Seguidores, Seguidos, Popularidade Relativa e Degree a melhor transformação foi a logarítmica e para as variáveis Clustering_N, Clustering_M e Eigenvector Centrality foi com a utilização da raiz quadrada.

Tabela 6– Desvio padrão, assimetria e curtose após o tratamento dos dados.

	Seguidores	Seguidos	Popul. Relativa	Degree	Cluste-ring_N	Cluste-ring_M	Eigenvector Centrality
Dados Originais							
Des.Padrão	2787,6	1410,1	33,9	36,0	0,0678	0,1937	0,0031
Assimetria	9,3	12,3	21,5	34,9	2,7	3,7	4,7
Curtose	121,5	237,5	547,7	1451,0	15,0	13,7	38,3
Transformação Log							
Des.Padrão	0,6433	0,4992	0,2733	0,2514	0,0258	0,0616	0,0013
Assimetria	-0,035	-0,047	4,207	4,433	2,120	3,377	4,583
Curtose	0,599	0,853	26,477	30,808	7,652	10,985	36,732
Transformação Raiz							
Des.Padrão	20,914	14,221	1,525	1,382	0,115	0,232	0,027
Assimetria	3,460	3,030	11,556	14,600	0,918	2,632	1,548
Curtose	18,638	19,239	172,746	306,184	1,195	6,081	3,195

A transformação para logarítmica reduziu a assimetria e a curtose das variáveis Seguidores (de 9,3 para -0.035 e de 121,5 para 0,0059), Seguidos (de 12,3 para -0,047 e de 237,5 para 0,853), Popularidade Relativa (de 21,5 para 4,207 e de 547,7 para 26,477) e Degree (de 34,9 para 4,443 e de 1451,0 para 30,808). Já a transformação para Raiz reduziu a assimetria e a curtose das variáveis Clustering_N (de 2,7 para 0,918 e de 15 para 1,195), Clustering_M (de 3,7 para 0,263 e 13,7 para 6,081) e Eigenvector Centrality (de 4,7 para 1,548 e de 38,3 para 3,195). Mesmo com a melhoria da assimetria proporcionada por estas transformações, muitos casos

continuaram sendo apontados como *outliers* pela ferramenta BoxPlot, especialmente os com poucos e com grande número de seguidores.

5.4 ANÁLISE DAS SUPOSIÇÕES ESTATÍSTICAS: NORMALIDADE, LINEARIDADE E COLINEARIDADE

Para complementar a análise dos dados, faz-se necessário verificar a normalidade, linearidade e colinearidade das informações obtidas. O processo de verificação está descrito a seguir.

Em consonância com a explicação de Pallant (2001), de que um caso de distribuição normal deve representarum desenho em forma de curva simétrica, este item foi verificado em todos os indicadores das variáveis do modelo. Para tal, utilizamos o teste de Kolmogorov-Smirnov. O mesmo Pallant (2001) aponta que, para encaixar-se no conceito de normalidade do teste mencionado, os valores do teste não podem ser inferiores a 0,05.

A tabela 7 mostra o resultado dos testes de normalidade Kolmogorov-Smirnov e Shapiro-Wilk utilizando o SPSS. Todas as variáveis podem ser consideradas não-normais.

Tabela 7– Resultados dos testes de normalidade das variáveis após a transformação dos dados.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
l_Seguidores	.032	2785	.000	.995	2785	.000
l_Seguidos	.033	2785	.000	.992	2785	.000
l_Popularidade_Relativa	.203	2785	.000	.645	2785	.000
l_Degree	.191	2785	.000	.648	2785	.000
r_Clustering_N	.085	2785	.000	.951	2785	.000
r_Clustering_M	.482	2785	.000	.458	2785	.000
r_Eigenvector_Centrality	.127	2785	.000	.861	2785	.000

a. Lilliefors Significante Correction

Para observar a configuração linear ou não-linear, foi feita a observação dos gráficos de dispersão, levando-se em consideração que a menor aleatoriedade e maior concentração dos dados em forma de linha central indicaria uma configuração linear. Não ficou configurada a linearidade entre as variáveis entre as variáveis.

A multicolinearidade ocorre quando dois ou mais previsores tem uma forte correlação entre si. A tabela 8 mostra as variáveis Degree e Popularidade Relativa tem correlação de Spearman acima de 0,80, que é o limite colocado por Field (2009) para que os previsores possam apresentar problemas de multicoliaridade em regressão múltipla.

Tabela 8– Valores dos coeficientes de correlação de Spearman.

	LOG Seguidores	LOG Seguidos	LOG Pop. Relativa	LOG Degree	RAIZ Clustering N	RAIZ Clustering M	RAIZ Eigenvector Centrality
LOG Seguidores	1,000	0,745	0,617	0,448	-0,425	0,264	0,604
LOG Seguidos	0,745	1,000	0,030	-0,098	-0,341	0,196	0,612
LOG Pop. Relativa	0,617	0,030	1,000	0,861	-0,163	0,178	0,251
LOG Degree	0,448	-0,098	0,861	1,000	-0,033	0,141	0,140
RAIZ Clustering_N	-0,425	-0,341	-0,163	-0,033	1,000	0,020	-0,172
RAIZ Clustering_M	0,264	0,196	0,178	0,141	0,020	1,000	0,369
RAIZ Eigenvector Centrality	0,604	0,612	0,251	0,140	-0,172	0,369	1,000

De acordo com Hair et. Al (2005, p.167), “correlações acima de 0,9 é a primeira indicação de colinearidade substancial”. Nenhuma variável apresencou colinearidade acima desse valor. As colinearidades consideradas por Pestana e Gageiro (2003) como altas (acima de 0,7) encontradas foram entre o LOG Degree e o LOG da Popularidade Relativa (0,861) e o LOG Seguidos e LOG Seguidores (0,745), as moderadas (entre 0,4 e 0,69) foramentre LOG Seguidores e RAIZ Eigenvector Centrality (0,604), LOG Seguidores e RAIZ Eigenvector Centrality (0,612), LOG Seguidores e LOG Popularidade Relativa (0,617), LOG Seguidores e

LOG Degree (0,448), LOG Seguidores e RAIZ Clustering (-0,425). Todas as outras relações apresentaram colinearidade baixa.

5.5 COMPARAÇÃO ENTRE CONTÁGIO E ESTRUTURA DE REDE: ANÁLISE DE REGRESSÃO LOGÍSTICA

Tendo como variável independente uma variável dicotômica, e percebendo não-normalidade nas variáveis independentes, com o objetivo de analisar o papel das variáveis de estrutura da rede individual dos usuários da rede social Twitter na difusão ou não das *hashtags* do estudo, seguiu-se a recomendação de Hair et. Al (2005) pela analise de regressão logística, pela robustez do teste estatístico e sua resistência às características de não normalidade. Os passos dessa analise, desde a seleção das variáveis, opção pela não-divisão da amostra, avaliações de ajuste, estimação do modelo, analises de significâncias estatística até os diagnósticos estão apresentados a seguir.

5.5.1 Objetivos, Projeto de Pesquisa e Suposições para a Analise de Regressão Logística.

O objetivo da presente analise alinha-se com o objetivo desta dissertação em analisar o papel das variáveis de estrutura da rede individual dos usuários da rede social Twitter na difusão ou não das *hashtags* estudadas. Para tal, foram selecionadas as variáveis independentes características de estrutura de rede social Degree, Clustering, Popularidade Relativa, Número de Seguidores, Número de Seguidos e Eigenvector Centrality e a variável dependente ContágioBin.

Já o tamanho da amostra atendeu o critério recomendado da proporção de Hair et. Al (2005) (de 20 para 1), com 2785 observações para 6 variáveis (464 para 1) na amostra da analise. Dada a robustez da técnica estatística para violação das suposições de igualdade das matrizes de variância/covariância, ela é recomendada por Hair Et. Al (2005) para aplicação na situação deste estudo.

5.5.2 Estimação do Modelo de Regressão Logística e Avaliação do Ajuste Geral

Em um primeiro momento, a fim de testar o modelo base dessa dissertação, um modelo de regressão logística foi estimado a partir do método de entrada forçada, com todas as variáveis consideradas nas hipóteses do trabalho. Foram estabelecidos o valor do logaritmo de verossimilhança e os pseudo-R^2.

Tabela 9 – Apresentação do -2LL e pseudo R^2

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
3051,303	,245	,328

O valor de -2LL encontrado foi 3051,30 e o pseudo-R^2 de Nagelkerke foi 0,328. Hair et. Al (2005) recomenda interpretar os valores de pseudo-R^2 com muita cautela, visto que esses valores não representam o R^2 como na regressão linear.

Tabela 10 – Apresentação dos índices de significância (Sig.) e taxas de previsão exp(B).

Variáveis	B	Wald	Sig.	Exp(B)
l_Seguidores	,346	,975	,323	1,413
l_Seguidos	,418	1,187	,276	1,519
l_Popularidade_Relativa	1,919	6,160	,013	6,811
l_Degree	-1,125	4,163	,041	,325
r_Clustering_N	2,896	41,494	,000	18,096
r_Eigenvector_Centrality	40,780	247,721	,000	5.13E+20
Constante	-4,329	88,927	,000	,013

Das variáveis testadas, a partir do método de entrada forçada do teste de Regressão Logística, as variáveis LOG Popularidade Relativa (0,13), LOG Degree (0,041), LOG Clustering-N (0,000) e LOG Eigenvector Centrality (0,000) apresentaram um índice de significância menor que 0,05, enquanto as variáveis LOG Seguidores (0,323) e LOG Seguidos (0,276) apresentaram índices de significância acima do critério de 0,05 proposto por Hair et. Al (2005).

Em um segundo momento, para estimar o melhor modelo de regressão possível a partir das variáveis independentes, realizou-se uma regressão logística no método Stepwise Forward a partir do critério do melhor índice Wald. Os passos dessa regressão estão apresentados a seguir.

Tabela 11 - Variáveis na equação

		B	Wald	Sig.	Exp(B)
Passo 1	r_Eigenvector_Centrality	47,275	455,352	,000	3.40E+23
	Constant	-1,621	460,835	,000	,198
Passo 2	l_Seguidores	,620	51,894	,000	1,859
	r_Eigenvector_Centrality	39,142	258,559	,000	9.98E+19
	Constant	-2,928	210,327	,000	,053
Passo 3	l_Seguidores	,846	76,223	,000	2,330
	r_Clustering_N	2,621	38,325	,000	13,743
	r_Eigenvector_Centrality	39,214	258,087	,000	1.07E+20
	Constant	-4,114	200,383	,000	,016

Dessa forma, as variáveis a entrar no modelo foram a Eigenvector Centrality (índice Wald de 455,352), Seguidores (índice Wald de 51,894) e Clustering N (índice Wald de 38,352). Após o terceiro passo, nenhuma outra variável apresentava grau de significância aceitável de acordo com os critérios do teste. A Tabela 11 apresenta as variáveis componentes do modelo e suas respectivas variações na proporção das probabilidades.

Os coeficientes estimados para as variáveis dependentes e a constante podem ser considerados com significância estatística a um nível de 0,001. A razão da desigualdade, gerada a partir das variáveis do modelo, pode ser expressado a partir da seguinte fórmula (onde EC = Eigenvector Centrality, Cl = Clustering e Se = Seguidores):

$$\frac{prob.contágio}{prob.\tilde{n}\ contágio} = e^{-4.114\ +\ 39{,}214.\sqrt{EC}+2{,}621.\sqrt{Cl}\ +0{,}846.log(Se)}$$

O teste Chi-quadrado para o valor de -2LL em relação ao modelo base a estabelece suporte para a aceitação do modelo. O teste estatístico da medida Hosmer e Lemeshow de ajuste geral indica que não houve qualquer diferença estatisticamente significante entre as classificações observadas e previstas, conforme apresentado na tabela 12.

Tabela 12 – Apresentação dos testes Chi-quadrado e de significância para a variação do valor -2LL em relação ao modelo base

Teste Hosmer e Lemeshow		
Passo	Chi-square	Sig.
1	58,995	,000
2	51,908	,000
3	44,389	,000

O teste qui-quadrado para a variação do valor -2LL em relação ao modelo base deu suporte para concluir a inclusão das 3 variáveis no modelo foi estatisticamente significante ao nível de 0,001. A medida de Hosmer e Lemeshow de ajuste geral indica que não existe diferença estatisticamente significante entre as classificações observadas e previstas. Com isso, o modelo logístico pode se considerado significante.

Para fins exploratórios, buscando um maior detalhamento do fenômeno estudado, em um terceiro momento de análise, também foram realizadas regressões logísticas (método Enter) das mesmas variáveis analisadas anteriormente para cada uma das *hashtags* componentes da amostra, conforme apresentado a seguir:

Tabela 13 - Sumário do modelo

Variáveis	-2 Log likelihood	Cox & Snell R Square
#quesevayalamafia	1,148,649	,362
#10pessoasquesonhariaemconhecer	1,081,429	,254
#letsmakeitawkward	683,111	,105

Para estas análises, foram obtidos diferentes índices de R^2 e -2LL, sendo que a *hashtag* #letsmakeitawkward teve menor desempenho em ambos (0,105 e 683,111), seguida da #10pessoasquesonhariaemconhecer (0,254 e 1.081,429). A hashtag com melhor desempenho encontrado foi a #quesevayalamafia (0,362 e 1.148,640).

Os resultados das regressões foram diferentes entre si, tanto em variáveis significantes quanto em taxas de influência de cada uma das variáveis no modelo.

Tabela 14 - Variáveis na equação

		B	Wald	Sig.	Exp(B)
#quesevayalamafia	l_Seguidores	,573	,772	,380	1,774
	l_Seguidos	-,737	1,100	,294	,478
	l_Popularidade_Relativa	-,442	,088	,766	,643
	l_Degree	-,880	,681	,409	,415
	r_Clustering_N	,363	,191	,662	1,438
	r_Eigenvector_Centrality	75,000	173,229	,000	3.73E+35
	Constant	-1,740	4,803	,028	,176
#10pessoasquesonhariaemconhecer	l_Seguidores	1,060	3,002	,083	2,886
	l_Seguidos	,044	,005	,946	1,045
	l_Popularidade_Relativa	1,219	1,067	,302	3,384
	l_Degree	-,815	1,108	,293	,443
	r_Clustering_N	4,733	32,575	,000	113,670
	r_Eigenvector_Centrality	36,541	69,880	,000	7.40E+18
	Constant	-5,327	56,869	,000	,005
#letsmakeitawkward	l_Seguidores	8,362	12,959	,000	4,279,120
	l_Seguidos	-7,625	10,848	,001	,000
	l_Popularidade_Relativa	-12,256	7,790	,005	,000
	l_Degree	-,297	,040	,842	,743
	r_Clustering_N	4,260	15,643	,000	70,785
	r_Eigenvector_Centrality	17,547	10,306	,001	4.18E+10
	Constant	,035	,000	,984	1,035

Para a difusão da *hashtag* #letsmakeitawkward, somente a variável Raiz Eigenvector Centrality (a nível de 0,001) de estrutura da rede foi significante,

enquanto para a hashtag #10pessoasquesonhariaemconhecer, puderam ser consideradas a Raiz Clustering (a nível de 0,001) e Raiz Eigenvector Centrality (a nível de 0,001) e para a #letsmakeitawkward, foram significantes a LOG Seguidores (0,001), LOG Seguidos (0,001), LOG Popularidade Relativa (0,005), RAIZ Clustering (0,001) e RAIZ Eigenvector Centrality (0,001).

Para todos os três casos, o teste Chi-quadrado para o valor de -2LL em relação ao modelo base a estabeleceu suporte para a aceitação do modelo. O teste estatístico da medida Hosmer e Lemeshow de ajuste geral indicou que não houve qualquer diferença estatisticamente significante entre as classificações observadas e previstas, conforme apresentado na tabela 15.

Tabela 15 - Teste Hosmer e Lemeshow

Variáveis	Chi-square
#quesevayalamafia	20,156
#10pessoasquesonhariaemconhecer	8,737
#letsmakeitawkward	6,432

5.6 RELAÇÃO ENTRE POTÊNCIA DE CONTÁGIO E ESTRUTURA DE REDE PARA OS USUÁRIOS QUE TRANSMITIRAM A HASHTAG: ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA

Sabendo-se que as variáveis componentes desse trabalho não atenderam os testes de normalidade recomendados por Hair et. Al (2005) e que deve haver ressalva na generalização dos resultados da regressão linear, o teste foi realizado a fim de explorar a relação do Clustering ao nível da mensagem (Clustering-M) na potência da difusão de informação.

Tabela 16 – Avaliação de melhora da previsão do modelo estimado

SUMÁRIO DO MODELO			
R Contágio = 1	R Quadrado	R Quadrado Ajustado	Erro Padrão
,610	,372	,368	,0112091193503

Para os casos em que a difusão gerou contágio (indivíduos que twittaram a *hashtag* e tiveram seguidores que também twittaram), o modelo estimado proporcionou melhora de 36,8% na previsão, com um erro padrão de 0,11.

Tabela 17 – Quoeficientes da Regressão Linear

COFICIENTES DA REGRESSÃO LINEAR				
	Unstandardized Coefficients		Standardized Coefficients	Sig.
	B	Std. Error	Beta	
(Constant)	,040	,004		,000
l_Seguidores	-,032	,003	-1,288	,000
l_Seguidos	,015	,003	,503	,000
l_Popularidade_Relativa	,037	,006	,799	,000
l_Degree	-,010	,004	-,201	,013
r_Clustering_N	-,003	,004	-,021	,474
r_Clustering_M	,010	,001	,228	,000
r_Eigenvector_Centrality	,085	,014	,176	,000

Todas as variáveis com excessão da RAIZ Clustering e LOG Degree foram significantes para a regressão a um nível 0,001 e o LOG Degree foi significante a um nível 0,02. As variáveis com maiores contribuições positivas relativas para a predição (coeficientes Beta positivos) foram LOG Popularidade Relativa (0,80), LOG Seguidos (0,50) e RAIZ Clustering-M (0,23). O LOG Seguidores, RAIZ Clustering-N (-0,21) e O LOG Degree (-0,20) tiveram as maiores contribuições negativas.

5.7 VERIFICAÇÃO DAS HIPÓTESES A PARTIR DA COLETA E DO TESTE

A partir da coleta dos dados e da analise de regressão logística, foi possível verificar as hipóteses propostas no modelo. Conforme apresentado a seguir, pôde-se confirmar seis das sete hipóteses da pesquisa:

H_1: haja correlação significativa positiva entre o grau de clustering de um determinado ponto de uma rede e a eficiência da transmissão de informação neste ponto.

O teste da regressão logística confirmou esta hipótese a um nível de significância de 0,005 para o conjunto total da amostra, no entanto, não se pôde confirmar a hipótese ao analisar a *hashtag* #quesevayalamafia.

H_2: haja correlação significativa positiva entre o grau de *clustering* de vizinhos adotantes de um determinado ponto de uma rede e a eficiência da transmissão de informação neste ponto.

O teste da regressão linear confirmou esta hipótese para o conjunto de usuários do twitter que emitiram a *hashtag* e obtiveram êxito no contágio a um nível de significância de 0,001 para o conjunto total da amostra.

H_3: a eficiência da transmissão de informação em um ponto da rede tenha correlação significativa positiva com a proporção entre a quantidade de conexões deste ponto da rede e a média de conexões de seu vizinho.

O teste da regressão logística confirmou esta hipótese a um nível de significância de 0,041 para o conjunto total da amostra, no entanto, não se pôde confirmar a hipótese ao analisar as *hashtags* separadamente.

H_4: Quanto maior a quantidade de conexões de acesso a informação de um ponto da rede, maior será a eficiência da transmissão de informação desse ponto.

O teste da regressão logística não confirmou esta hipótese (nível de significância de 0,276) para o conjunto total da amostra, mas confirmou na analise individual da *hashtag* #letsmakeitawkward (a um nível de significância de 0,001).

H_5: Quanto maior a quantidade de conexões de envio de informação de um ponto da rede, maior será a eficiência da transmissão de informação desse ponto.

O teste da regressão logística não confirmou esta hipótese (nível de significância de 0,323) para o conjunto total da amostra, mas confirmou na analise individual da *hashtag* #letsmakeitawkward (a um nível de significância de 0,001).

H_6: Quanto maior a proporção entre a quantidade de conexões de envio de informações e a quantidade de conexões de recebimento de informações de um ponto da rede, maior será a eficiência da transmissão de informação desse ponto.

O teste da regressão logística confirmou esta hipótese (nível de significância de 0,013) para o conjunto total da amostra, mas confirmou na analise individual da *hashtag* #letsmakeitawkward (a um nível de significância de 0,005).

H_7: quanto maior a *eigencevtor centrality* de um ponto, maior será a eficiência da transmissão da informação deste ponto.

O teste da regressão logística confirmou esta hipótese (nível de significância de 0,001) para o conjunto total da amostra, como também confirmou na análise individual da de cada *hashtag* (a um nível de significância de 0,001).

6 CONCLUSÕES, RESTRIÇÕES E SUGESTÕES

A finalidade deste capítulo é apresentar as conclusões do projeto considerando os objetivos anteriormente propostos e o caminho percorrido para alcançá-los. Além disso, também são apresentadas limitações identificadas neste projeto e apresentadas sugestões de novos estudos futuros a partir da continuidade ou do aprofundamento das construções abordadas neste trabalho.

6.1 CONCLUSÕES

A compreensão dos fenômenos da “viralização” (sobretudo a difusão de informação em rede social online) e a identificação de caminhos possíveis para a transmissão de uma mensagem nos contextos de redes sociais, que têm sido de grande interesse dos estudiosos e administradores do marketing devido à popularização e crescimento das plataformas de redes sociais online, e da carência de comprovações científicas e práticas das teorias de estrutura de redes sociais desenvolvidas pela sociologia, antropologia e matemática em estruturas reais, o projeto em estudo teve como objetivo compreender o papel da estrutura da rede social online no processo de difusão de informação.

Para tal, foram abordadas hipóteses teóricas de estrutura de rede e difusão de informação para a construção de um modelo que poderia oferecer alguns apontamentos de características estruturais de redes sociais capazes de influenciar a difusão de mensagens entre os usuários.

Sendo assim, o estudo utilizou computadores de altíssimo desempenho para coletar e analisar 846.441 usuários da plataforma de rede social *Twitter*.com, interligados por laços de transmissão de informação, totalizando 2790 emissores de três *hashtags* específicas. Ressaltada a ausência de intenção de generalização dos resultados do trabalho, vistos os métodos adotados e a população amostra utilizadas, as conclusões a serem apresentadas contribuem para a discussão das teorias de difusão de informação, teorias de redes sociais e para uma visão mercadológica, sociológica e antropológica mais ampla do contexto das redes sociais online e do fenômeno da “viralização” de mensagens, produtos, conceitos e comportamentos.

As conclusões referentes aos objetivos específicos do projeto (1), às contribuições teóricas (2) e as contribuições gerenciais (3) do projeto são apresentadas nas próximas três partes do trabalho.

6.2 CONCLUSÕES DOS OBJETIVOS PROPOSTOS

Os objetivos propostos no estudo foram contemplados a partir da analise de variáveis indicadoras de característica de rede e de difusão de *hashtags* no Twitter e verificaram o papel da estrutura da rede social online no processo de difusão de informação especificamente analisando três relações:

(1) Verificou-se a relação entre quantidade de conexões, a proporção seguidores/seguidos de um membro da rede e a difusão de informação neste ponto específico da rede social.

Para a amostra do estudo, pode-se concluir que tanto as variáveis da quantidade de conexões (variáveis Seguidores e Seguidos) quanto a proporção seguidores/seguidos (variável Popularidade Relativa) apresentaram significância estatística (a nível de 0,001) na explicação do fenômeno da difusão (Variável ContágioBin).

(2) Verificou-se a relação entre a quantidade de interconexões entre os vizinhos diretos de um membro de uma rede social online e a difusão de informação neste ponto específico.

Para a amostra do estudo, pode-se concluir que a variável representativa da quantidade de interconexões entre os vizinhos diretos de um membro da rede (variável Clustering_N) apresenta significância estatística (a nível de 0,005) na explicação do fenômeno da difusão (variável ContágioBin).

(3) Verificou-se a relação entre o grau a que um determinado membro de uma rede social online localiza-se no menor caminho entre outros membros da rede e a difusão de informação neste ponto específico da rede.

Para a amostra do estudo, pôde-se concluir que a variável representativa do grau a que um determinado membro de uma rede social localiza-se no menor caminho entre outros membros, a nível da proporção entre a média de seguidores dos membros e a média de seguidores do usuário em análise (variável Degree),

apresenta significância estatística (a nível de 0,001) na exploração do fenômeno da difusão (variável ContágioBin). No entanto, devido à ausência de capacidade computacional para calcular a variável Betweeness, não foi possível verificar a significância estatística do indicador de menor caminho a nível da proporção entre a quantidade de menores caminhos possíveis e a real presença desses caminhos.

6.3 CONTRIBUIÇÕES TEÓRICAS

Ao longo deste trabalho, foi possível, além das conclusões provenientes dos objetivos específicos, contribuir à teoria de acordo com o que está apresentado a seguir.

Ao nível da estrutura de rede, a proposta de Katona, Zubcsek e Miklos (2011), de que se pode identificar potenciais “atores” influenciadores, como também obter informações sobre o fluxo de uma informação na rede a partir da observação da estrutura da rede foi extrapolada para a amostra do Twitter, ao nível da difusão individual de *hashtags*. A sugestão de Valente e Roger (1995) e Ryan e Gross (1943), de que, além de característica estritamente econômicas, fatores sociais também são responsáveis pela difusão foi validada para a amostra selecionada, sob o aspecto das características da estrutura das redes de comunicação.

Ao nível das características de redes sociais, esta dissertação contribuiu extrapolando indicadores das teorias sociais de Freeman (1977), French e Raven (1960), Granovetter (1973), Burt (1992), Watts e Strogatz (1998) e Moody (2009) para o contexto das Redes Sociais Online reais e de larga escala. As significâncias encontradas para as variáveis no teste da regressão podem contribuir teoricamente na analise das teorias de rede conforme especificados a seguir:

- A explicação da difusão de *hashtag*s no twitter a partir da Teoria da Força dos Laços Fracos ganhou corpo a partir da confirmação da significância do indicador de Clustering no contagio, reduzindo assim a expectativa da influência do efeito negativo na eficiência da transmissão previsto por Moody (2009).

- A Teoria do Buraco Estrutural com a finalidade de explicar a difusão teve seu peso confirmado no Twitter com a confirmação da significância do Degree no Contágio.
- O apontamento, por Liu, Madhavan e Sudharshan (2005), da relação inovador/imitador como elemento de influência da difusão em redes sociais, bem como a percepção de Lin (1999), de que elementos de rede com maior conexões de saída de informação têm melhor desempenho foram extrapoladas para a rede social Twitter.

Além das contribuições pontuais de confirmação e extrapolação das teorias inerentes à difusão de informação e à analise de estrutura de redes sociais, o presente estudo contribuiu à teoria com a sugestão de um modelo inicial de previsão logístico do contágio na rede Twitter, composto por apenas 3 indicadores de estrutura de rede:

$$\frac{prob.contágio}{prob.\tilde{n}\ contágio} = e^{-4.114 + 39{,}214.\sqrt{EC} + 2{,}214.\sqrt{Cl} + 0{,}846.log(Se)}$$

6.4 CONTRIBUIÇÕES METODOLÓGICAS

Ao que tange o método da coleta de dados, o presente trabalho abre um novo campo de estudo das redes sociais, a partir de uma ferramenta de coleta de dados em larga escala, com alta eficiência. O método se comprovou eficiente em reproduzir a estrutura de redes sociais altamente complexas e transformá-las em variáveis passíveis de analises quantitativas.

6.5 CONTRIBUIÇÕES GERENCIAIS

Tomadas as conclusões do estudo como base, algumas decisões gerenciais podem ser adotadas a fim de melhor aproveitar, estimular, reter e acelerar o processo de difusão de informações em redes sociais online. Essas decisões podem reduzir as chances de falha de determinados produtos ou ainda prever e acurar

tomadas de decisões mercadológicas com relação a lançamentos de produtos em ambientes de redes sociais online.

A comprovação da eficiência e indicação de elementos estruturais de redes sociais indicativos de melhores chances de difusão de informação podem oferecer ferramentas aos profissionais de administração e marketing para tomadas de decisão com relação a estratégias em inúmeras esferas, desde a comunicação até o desenvolvimento de produto. As comprovações desta dissertação podem servir de auxílio para:

- Seleção de personagens influenciadores: para ser utilizados em estratégias de difusão de produtos ou lançamentos, a fim de obter maior eficiência.
- Previsão de eficiência de personagens influenciadores: no nível da avaliação e comparação de atores para a seleção da estratégia.
- Seleção de caminhos de difusão de uma informação: observando o panorama geral da rede, traçando estratégias de caminhos por onde a difusão possa ocorrer com maior eficiência possível.
- Previsão de eficiência de caminhos de difusão de uma informação: no nível da avaliação e comparação entre diferentes caminhos já mapeados.
- Avaliação das chances de falha de uma difusão: gerando um cálculo de risco das estratégias a partir da estrutura da rede.
- Comparação com outras estratégias: comparação entre riscos e chances de sucesso de diferentes estratégias, a partir da estrutura da rede.
- Previsão de eficiência de estratégias de difusão.

6.6 RESTRIÇÕES DO ESTUDO REALIZADO

Mesmo tendo sido coerentemente escolhidas para alcançar o objetivo do estudo, algumas limitações pontuais podem ser apresentadas.

- O levantamento realizado para a pesquisa, apesar de atender os objetivos propostos, limita-se à descrição superficial da amostra selecionada, não identificando causas ou explicações com relação às relações encontradas.

- As conclusões do estudo são limitadas apenas à descrição do fenômeno na rede social Twitter e às *hashtags* analisadas, no período de tempo da coleta.
- As conclusões obtidas a partir de apenas três *hashtags* pesquisadas carecem de verificações em outras *hashtags*, haja visto que houve grande variação de comportamento da estrutura da rede, mesmo entre as três *hashtags* analisadas.
- Apesar das variáveis do estudo terem sido construídas e selecionadas a partir da teoria, deve-se levar em consideração a existência de outras variáveis que não foram incluídas na pesquisa ou da existência de melhores formas de representar os fenômenos descritos na teoria.
- A seleção da unidade de caso como os usuários que realizaram a emissão das *hashtags*, apesar de adequada para o estudo da influência da estrutura da rede na difusão, limitou-se ao estudo da difusão individual, não avaliando a evolução da difusão por mais de 1 relação de difusão.
- O método de regressão logística, apesar de robusto para as características dos dados do estudo, não contemplou intensidades de contágio, considerando a mesma eficiência para todos os casos que obtiveram sucesso na difusão da informação.
- O método de regressão linear utilizado na pesquisa não atendeu às premissas de normalidade das variáveis, não sendo, assim, passível de generalização.
- O método de coleta de dados desenvolvido ainda carece de verificações em outras amostras e redes sociais, para que possa alcançar generalidade.

6.7 SUGESTÕES PARA FUTURAS PESQUISAS

Vista a grande relevância da difusão de informação em rede social, a carência de estudos e comprovações em redes sociais reais e a incapacidade de generalização das conclusões desse trabalho, com a intenção de aumentar a

compreensão sobre o tema abordado, algumas sugestões de estudos futuros fazem-se necessárias e podem ser propostas.

- Faz-se necessária a avaliação deste mesmo fenômeno em diferentes contextos de redes sociais, com diferentes *hashtags* e diferentes espaços de tempo.
- Faz-se necessária a replicação do método em outras pesquisas, para verificação de consistência e replicabilidade.
- Faz-se necessária melhor explicação dos fenômenos inerentes à estrutura de rede e difusão, sobretudo em *hashtags* específicas.
- Para melhor explicar o fenômeno, faz-se necessária a busca de outras variáveis componentes do modelo ou de ajustes nas variáveis atuais do modelo.
- Estudos envolvendo o acompanhamento da evolução da difusão da informação através dos caminhos da rede poderiam oferecer uma melhor explicação das teorias de difusão de informação.
- A inclusão da “potência” de difusão no modelo de regressão linear com variáveis normais poderia trazer melhores elucidações no que tange a influência da estrutura das redes na difusão.

REFERÊNCIAS BIBLIOGRÁFICAS

ALDERSON, Arthur S.; Jason Beckfield. **Power and Position in the World City System**. American Journal of Sociology, 109:811-851, 2004.

AKINOLA, A. **An application of the Bass model in the analysis of diffusion of cocospraying chemicals among Nigerian cocoa farmers**. **Journal** of Agricultural Economics, 37 (3), p.395-404, 1986.

ANGLES, Renzo; GUTIERREZ, Claudio. **Survey of graph database models**. ACM Comput. Surv.40, 1, Artigo 1, p. 1-39, fev. 2008. Disponível em: http://doi.acm.org/10.1145/1322432.1322433>. Acesso em:

BAILEY, N.T.J. **The Mathematical Theory of Infectious Diseases and its Applications.** Charles Griffen. Londres. 1975.

BASS, Frank M. **A new product growth model for consumer durables**. Management Sci. 15, p. 215-227. 1969.

Beevolve (2012), Disponível em: <http://www.beevolve.com/*twitter*-statistics/>

BONACICH, P. **Factoring and weighting approaches to clique identification**. Journal of Mathematical Sociology 2, p. 113–120, 1972.

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

BORGATTI, S. P.; HALGIN, D. S. On Network Theory. **Organization Science. Informs**. p. 1–14, 2011.

BREIGER, R. L. **The analysis of social networks**. In M. Hardy & A. Bryman (Eds.), Handbook of data analysis, Londres, 2004.

BROWN, T. E.; Ulijn, J. M. **Innovation, entrepreneurship and culture: The interaction between technology, progress and economic growth**. MA: Northampton, 1981.

BULTE, C. Van den; JOSHI, Y.V. **New product diffusion with influentials and imitators**. Marketing Science 26, p.400–421, 2006.

BURT, R. S. **Structural holes: The social structure of competition**. MA: Harvard University Press, Cambridge, 1992.

BURT, R. S. **Social Contagion and Innovation: Cohesion Versus Structural Equivalence.** American Journal of Sociology 92(6): p.1287-1335, 1987.

BURT, R. S. **Brokerage and Closure**. Chapter 3: Closure, trust and reputation. Oxford University Press, p.93–166, 2005.

BURT, R. S. **Structural Holes and Good Ideas**. American Journal of Sociology v.110, p.349-399, 2004.

BURT, R. S. **The network structure of social capital**. Em R. I. Sutton e B. M. Staw (Eds.), Research in organizational behavior. Greenwich, CT: JAI Press, 2000.

BURT, R. S. **Social origins of good ideas**, 2002. Disponível em: <http://www.analytictech.com/mb709/readings/burt_SOGI.pdf>. Acesso em: 20 mai. 2012.

CHANDRASEKARAN, D., & Tellis, G. J. **Getting a grip on the saddle: Cycles, chasms, or cascades?** PDMA Research Forum, p.21–22, Atlanta, out. 2006.

CHANDRASEKARAN, D., & Tellis, G. J. (2007). **A critical review of marketing research on diffusion of new products**. In N. K. Malhorta (Ed.), Review of marketing research, p. 39−80, Armonk, NY: M. E. Sharpe.

CODD, E. F. **A relational model of data for large shared data banks**. Commun. ACM 13, 6, p.377-387, 1970.

CODD, E. F. **Data models in database management**. In Proceedings of the 1980 Workshop on Dataabstraction, Databases, and Conceptual Modeling. ACM Press, p.112–114, 1980.

COLEMAN, James S. (1988), Social **Capital in the Creation of Human Capital**. American Journal of Sociology, 94, 95–120.

COLEMAN, J. S. **The foundations of social theory.** Harvard University Press, Cambridge, 1990.

COLEMAN, J.S.; E. Katz; H. Menzel. **Medical Innovation: A Diffusion Study**. New York, 1966.

COOLEY, C. H. **Human Nature and the Social Order**. New York: Schocken, 1964.

CRESWELL, John W. **Projeto de Pesquisa: métodos qualitativos, quantitativo e misto**. Ed. 3, São Paulo, 2010.

DURKHEIM, E.; MONTESQUIEU; ROUSSEAU. **Forerunners of Sociology.** MI: University of Michigan Press, 1965.

EBIZMBA RANK: Disponível em: <http://www.ebizmba.com/articles/social-networking-websites>. Acesso em: 23 mai. 2012, 16:00.

FOURT, L.; WOODLOCK , Joseph. **Early Prediction of Market Success of New Grocery Products**. Journal of Marketing, 25(2), p.31–38, 1960.

FRENCH, John R. P. Jr.; BERTRAM H. Raven. **The bases of social power.** Group dynamics, D. Cartwright and A. Zander (eds.), chap. New York: Harper and Row, p.607–623, 1960.

GATIGNON, Hubert; ELIASHBERG, Jehoshua; ROBERTSON, Thomas S. **Modeling Multinational Diffusion Patterns: An Efficient Methodology**. Marketing Science, 8 (3), p.231–247, 1989.

GNEISER, M.; HEIDEMANN, J.; LANDHERR, A.; KLIER, M.; PROBST, F. **Valuation of online social networks taking into account users' inter- connectedness.** Appears in: ISeBM Special Issue, 2010.

GODIN, Seth. **Permission Marketing: Turning Strangers Into Friends And Friends Into Customers**. Simon e Schuster, ed. 1, mai. 1999.

GOLDENBERG, Jacob; LEHMANN, Donald R., SHIDLOVSKI, Daniella et al. **The role of expert versus social opinion leaders in new product adoption**. Marketing Science Institute, Cambridge, MA, p.06-124, 2006.

GRANOVETTER, M. **The strength of weak ties**. American Journal of Sociology, vol. 78, p.1360-80, 1973.

GRANOVETTER, M. **Getting a Job: A Study of Contacts and Careers**. Cambridge : Harvard University Press, 1974.

GRANOVETTER, M. **Economic action and social embeddedness.** American Journal of Sociology, vol. 91, p. 481-510, 1985.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 5.ed. São Paulo: Editora Atlas, 1999.

HAGERSTRAND, Torsten. **Innovation Diffusion as a Spatial Process.** Chicago: University of Chicago Press, 1953.

HAHN, Minhi; PARK, Sehoon; KRISHNAMURTHI, Lakshman; ZOLTNERS, Andris. **Analysis of New Product Diffusion Using a Four Segment Trial-Repeat Model**. Marketing Science 13 (3), p.224–247, 1994.

HAIR JR., J. F.; ANDERSON, R.E., TATHAN, R.L., BLACK, W.C. **Análise multivariada de dados**. 5 Ed. Porto Alegre: Bookman, 2005.

HAWKINS, DI; BEST, R J e CONEY, K. A. **Consumer Behaviour: Implications for Marketing Strategy**. Homewookd, Boston: PBI/IR WIN, 1989.

HILL, Shawndra; PROVOST, Foster; VOLINSKY, Chris. **Network-based marketing: Identifying likely adopters via consumer networks**. Statistical Science 22(2), 2006.

HUAN, Jun; WANG, Wei; PRINS, Jan; YANG, Jiong. **SPIN: mining maximal frequent subgraphs from graph databases**. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04). ACM, New York, NY, USA, p.581-586, 2004. Disponível em: http://doi.acm.org/10.1145/1014052.1014123.

JEULAND, A. **Empirical Generalisations in Marketing**. Draft Proceedings, SEI Center for Advanced Studies in Management. The Wharton School of the University of Pennsylvania, p.16-18, fev. 1990.

KERLINGER, F. N. **Metodologia da pesquisa em ciências sociais: um tratamento conceitual**. 5.ed. São Paulo: EDUSP, 1980.

KATONA, Zsolt; ZUBCSEK, Peter Pal; SARVARY, Miklos. **Network Effects and Personal Influences: The Diffusion of an Online Social Network**. Journal of Marketing Research: vol. 48, n. 3, p.425-443, 2011.

KATZ, N. LAZER, D; ARROW, Holly; CONTRACTOR, Noshir. **Network Theory and Small Groups.** Small Group Research, vol. 35. n. 3, p.307-332, jun. 2004.

KEENAN, Andrew; SHIRI, Ali. **Sociability and social interaction on social networking websites.** Library Review, vol. 58, p.438 – 450, 2009.

KEMPE, D.; KLEINBERG, J.; TARDOS, E. **Maximizing the spread of influence through a social network**. In KDD ’03: Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge discovery and data, p. 137–146, 2003.

KERLINGER, F. **Metodologia da pesquisa em ciências sociais**. 5a edição. São Paulo: EPU, 1979.

KILDUFF, M. **Serendipity vs. strategy: A tale of two theories**. Pre- sentation, Intra-Organizational Networks Conference, University of Kentucky, Lexington, abr. 2010.

KIRSCHBAUM, C. **Renascença da Indústria Brasileira de Filmes: Destinos Entrelaçados?** RAE, vol. 45. n 3. p. 58-71, 2006.

KISS, Christine; BICHLER, Martin. **Identification of influencers — Measuring influence in customer networks.** Decision Support Systems, v. 46, p. 233-253, dez. 2008.

KUNST, Laurien; KRATZER, Jan. **Diffusion of innovations through social networks of children**. Young Consumers: Insight and Ideas for Responsible Marketers, vol. 8, p.36 – 51, 2007.

LIBAI, Barack; BOLTON, Ruth; BÜGEL, Marnix; RUYTER, Ko. **Customer-to-Customer Interactions: Broadening the Scope of Word of Mouth Research**. Journal of Service Research, 3 ed., p. 267 – 282, 2010.

LIPMAN http://www.forbes.com/sites/victorlipman/2013/05/01/the-worlds-most-active-twitter-country-hint-its-citizens-cant-use-twitter/

LIU, Ben Shaw-Ching; MADHAVAN, Ravindranath; SUDHARSHAN, D. **DiffuNET: The impact of network structure on diffusion of innovation**. European Journal of Innovation Management, vol. 8, p.240 – 262, 2005.

MAHAJAN, Vijay; MULLER, Eitan; BASS, Frank M. **New-product diffusion models**. J. Eliashberg, G.L. Lilien, eds. Marketing (Handbooks in Operations Research and Management Science, Vol. 5). North-Holland, Amsterdam, Netherlands, p.349-408, 1993.

MARCONI, Maria de Andrade, LAKATOS, Eva Maria. **Fundamentos da metodologia cientifica.** 7. ed. São Paulo, Atlas, 2010.

MARTINO, F., & SPOTO, A. **Social Network Analysis : A brief theoretical review and further perspectives in the study of Information Technology**. Social Networks, 4(1), p.53 – 86, 2006.

MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. **Measurement and analysis of online social networks**. Proc of the 7th ACM SIGCOMM conf on internet measurement, San Diego, 2007.

MOODY, James; WHITE, Douglas R. **Social Cohesion and Embeddedness: A Hierarchical Conception of Social Groups**. American Sociological Review 68:103-27, 2003.

MOODY, James. **Network Structure and Diffusion**. Duke Population Research Institute. Disponível em: <http://papers.ccpr.ucla.edu/papers/PWP-DUKE-2009-004/PWP-DUKE-2009-004.pdf>. Acesso: 26 mar. 2012.

MCPHERSON, J. M.; SMITH-LOVIN, L. J.; COOK M. 2001. **Birds of a feather: Homophily in social networks**. Annual Rev. Soc. 27 415–444.

PETERSON, Robert A.; MERINO, Maria C. **Consumer Information Search Behavior and the Internet**. The University of Texas, 2003.

Pestana, M.; Gageiro, J. **Análise de dados para ciências sociais** – A complementaridade do SPSS. Lisboa: Edições Sílabo, 2003.

Scott, J. **Social Network Analysis**. Sage, Londres, 1991.

TELLIS, G. J. **A Critical Review of Marketing Research on Diffusion of New Products, in Naresh K. Malhotra.** Review of Marketing Research, vol. 3, p.39-80, Emerald Group Publishing Limited, 2007.

RYAN, B.; GROSS, N. C.. **The Diffusion of Hybrid Seed Corn in TWO Iowa Communities**. Rural Sociology, p. 15-24, mar. 1943.

ROBERTS (2012), http://gigaom.com/2012/10/10/the-typical-twitter-user-is-a-young-woman-with-an-iphone-and-208-followers/

ROGERS, E. M. (1983). Diffusion of innovations. 3 ed. New York: Free Press, 1983.

Rogers, E. M. Diffusion of innovations. New York: Free Press, 1962.

Rogers, E. M. Diffusion of innovations . 5 ed. New York: Free Press, 2003. [2]

STEPHEN, Andrew T.; DOVER, Yaniv; GOLDENBERG, Jacob. **You Snooze You Lose: Compar- ing the Roles of High Activity and Connectivity in Information Dissemination Over Online Social Networks, working paper**. INSEAD, França, 2010.

Semiocast (2013) Disponível em: http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US.

SCHROEDER(2013), *By the numbers, a few amazing twitter stats.* Disponível em: http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-*twitter*-stats/

TWITTER BLOG (2012), *Changes coming in version 1.1 of the Twitter API.* disponível em https://dev.twitter.com/blog/changes-coming-to-twitter-api

NEO4J (2013) *What is Neo4j? .* Disponível em: http://www.neo4j.org/learn/neo4j

TWITTER DEV BLOG (2013) https://dev.twitter.com/blog/changes-coming-to-twitter-api

TELLIS, G. J. Tellis. **A Critical Review of Marketing Research on Diffusion of New Products**, in Naresh K. Malhotra. Review of Marketing Research, vol. 3, p.39-80, Emerald Group Publishing Limited, 2007.

VALENTE, T.W. et al. **Social Network Associations with Contraceptive Use Among Cameroonian Women in Voluntary Associations**. Social Science and Medicine (45), p. 677-687, 1997.

VALENTE, T.W. **Network models of the diffusion of innovations**. Quantitative methods in communication. Cresskill, N.J.: Hampton Press. p. 171, 1995

VALENTE, T.W.; DAVIS, R.L. **Accelerating the Diffusion of Innovations Using Opinion Leaders**. The ANNALS of the American Academy of Political and Social Science, 566(1): p. 55-67, 1999.

VALENTE, T.W. **Network Models of the Diffusion of Innovations.** Cresskill: Hampton Press, 1995.

VALENTE, T.W.; ROGERS, E.M. **The origins and development of the diffusion of innovations: paradigm as an example of scientific growth.** Science Communication, vol. 1, 1995.

WASSERMANAND, S.; FAUST, K. **Social Network Analysis: Methodsand Applications**. Cambridge University Press, 1994.

WATTSAND, D.J.; STROGATZ, S.H. **Collective dynamics of'small-world networks. Nature**, 393:440–442, 1998.

WEJNERT, B. **Integrating models of diffusion of innovations: a conceptual framework.** Annual Review of Sociology, vol. 28, p. 297-326, 2002.

WELLMAN, B. **The community question: The intimate networks of East Yorkers**. American Journal of Sociology 84(5), p.1201–1231, 1979.

WRIGHT, M; CHARLETT, D. **New Product Diffusion in Marketing: An assessment of Two Approaches**. Marketing Bulletin, 6, art. 4, p. 32-41, 1995