ISCTE IUL
Instituto Universitário de Lisboa
Departamento de Ciências e Tecnologias de Informação

FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA
Departamento de Informática

Mecanismos de Popularidade e Difusão de Informação em Redes Sociais

António Jorge Filipe Fonseca

Tese especialmente elaborada para a obtenção do grau de

Doutor em Ciências da Complexidade

Orientada pelo Prof. Jorge Manuel Anacleto Louçã

Novembro, 2014

Júri da Prova

Presidente:

- Doutor Luís Antero Reto, Professor Catedrático do ISCTE-IUL

Vogais:

- Doutor Helder Manuel Ferreira Coelho, Professor Catedrático da Universidade de Lisboa no Departamento de Informática da Faculdade de Ciências
- Doutor Ernesto Costa, Professor Catedrático da Universidade de Coimbra no Departamento de Engenharia Informática
- Doutor Luís Alberto dos Santos Antunes, Professor Auxiliar da Universidade de Lisboa no Departamento de Informática da Faculdade de Ciências
- Doutor Pedro Gonçalves Lind, "Wissenschaftlicher Mitarbeiter", investigador, relator na Universidade de Oldenburg
- Doutor António Firmino da Costa, Professor Catedrático do ISCTE-IUL no Departamento de Sociologia

Orientação:

- Doutor Jorge Manuel Anacleto Louçã, Professor Auxiliar do ISCTE-IUL no Departamento de Ciências e Tecnologias da Informação

Agradecimentos

Gostaria de expressar os meus sinceros agradecimentos a todos os que me acompanharam na jornada da elaboração desta tese. O trabalho que envolveu foi acompanhado com expectativa por família, amigos e colegas.

Gostaria de expressar os meus agradecimentos ao meu orientador, o Professor Jorge Louçã, pelo cuidadoso apoio que me deu na redação da tese e no incentivo ao trabalho de investigação na área dos sistemas complexos.

Gostaria também de expressar os meus agradecimentos aos colegas de trabalho pela tolerância que me deram nas horas mais apertadas da laboração quotidiana a completar esta tese.

Finalmente gostaria de expressar a minha estima à minha família e aos meus pais pelo seu amor incondicional.

Abstract

This research investigates the mechanisms of formation of popularity in society, assuming that popularity is generated though processes of information diffusion.

A static model of the distribution of popularity by various entities is here proposed and validated. It is demonstrated that it fits a probabilistic distribution of exponential growth of popularity. Complementarily, two dynamic models are proposed, representing the evolution of popularity. The first model, named Ramification Model, is an exogenous impact model tracing the profile of the typical evolution of popularity triggered by a single external event. The second one, called Epidemy Model, represents the process of popularity formation when arising from internal dissemination of messages within a community. All models are validated with experimental data.

A case study, concerning communication data collected during the 2011 elections in Portugal, allowed measuring the influence of popularity, generated through Social Communication, on opinion dynamics. Two sociophysics opinion dymamics models, based on the Brownian Model of Infuence, and on the Social Impact Theory, were used to represent theoreticaly and quantitatively the dynamics of public debate in this period.

One of the most relevant results of the research concerns the understanding that the long term increasing of some entity's popularity, as a result of comunication processes between individuals, is independent from the entity subjective qualities, and it depends mainly from the comunication processes being used.

Keywords: Popularity, Information Difusion, Social Networks, Growth Models, Multi-agent Models, Opinion Dynamics, Election Prediction, Sociophysics

Resumo

Este trabalho de investigação estuda os mecanismos de formação de popularidade na sociedade, supondo que a popularidade é gerada por processos de difusão de informação.

Um modelo estático de distribuição de popularidade por diversas entidades é proposto e validado. Demonstra-se que este modelo se adequa a uma distribuição probabilística de crescimento exponencial da popularidade. Complementarmente, dois modelos dinâmicos para representar a evolução da popularidade são propostos. O primeiro, o Modelo de Ramificação, é um modelo exógeno de impacto que traça o perfil típico da evolução da popularidade a partir de um evento singular externo. O segundo, denominado Modelo de Epidemia, representa o processo de criação de popularidade quando esta surge da difusão interna de mensagens no seio de uma comunidade. Todos os modelos são validados com dados experimentais.

A magnitude da influência da popularidade na opinião pública, transmitida pela Comunicação Social, é particularmente analisada no estudo do caso das eleições presidenciais e legislativas, em 2011 em Portugal. Dois modelos sociofísicos multi-agente de dinâmica de opiniões, baseados no Modelo Browniano de Influência e na Teoria do Impacto Social, são usados para representar a dinâmica do debate político ocorrido neste período, em termos teóricos e quantitativos.

Entre os resultados relevantes dos trabalhos de tese, destaca-se a conclusão de que o crescimento a longo termo da popularidade de uma entidade, fruto dos processos de comunicação entre indivíduos, é independente das qualidades subjectivas daquela ou destes, dependendo sobretudo dos processos em que é comunicada.

Palavras-Chave: Popularidade, Difusão de Informação, Redes Sociais, Modelos de Crescimento, Modelos Multi-Agente, Dinâmica de Opinião, Previsão do Voto, Sociofísica

Índice

Lista de Figuras

Parte I

Introdução

Capítulo 1

Introdução

1.1 Âmbito da investigação

A popularidade é um fenómeno social que determina que alguém ou alguma coisa é preferido por um grande número de pessoas, que está "na moda". Em 2001 o filósofo Daniel Dennett comparou a consciência humana à 'fama no cérebro' Dennett [2005], sugerindo que a mente age, até um certo ponto, como uma câmara de eco onde agentes semi-autónomos se mobilizam para construir a consciência. Segundo este filósofo o ser humano está consciente de conteúdos que conseguem monopolizar por instantes os recursos corticais, em competição com muitas outras representações. Como a consciência, o fenómeno da popularidade na sociedade humana é um fenómeno importante porque determina quer processos económicos, quer processos políticos, quer ainda processos culturais que constroem a memória coletiva. A popularidade determina escolhas de consumo, atos eleitorais e a formação e evolução da opinião pública. Na génese da popularidade há características intrínsecas aos sujeitos. No entanto, como pretendemos demonstrar nesta tese, estas características não são decisivas na magnitude da popularidade. Como nota Dennett, consideremos o lançamento de um livro num hipotético programa televisivo de grande audiência. Suponhamos que no dia da transmissão acontece um acidente natural que distrai a atenção do público. O autor e o livro são os mesmos, antes e depois do programa, no entanto a sua popularidade fica gravemente afetada com a falta de atenção do público.

Numa época de conexão global como aquela em que vivemos, em que as pessoas se expõem cada vez mais através de atos comunicativos em redes sociais, a título individual ou em grupo, o problema da saliência no mundo mediático tornou-se relevante. A questão, que a nós próprios colocámos e que pretendemos responder nesta tese, deriva da compreensão do fenómeno da popularidade. Em cada minuto são gerados cerca de 350.000 *tweets* [1], são partilhadas 100 horas de vídeo no Youtube [2] e há 590.000 ações no Facebook [3]. A popularidade dos conteúdos na Internet, quer sejam informativos, de entretenimento ou como oportunidade de negócio para criadores e empresas de marketing, é um tema importante no mundo moderno. A facilidade com que estes conteúdos são criados propicia um dilúvio de informação. No entanto, no eco-sistema *online*, a lógica de que o vencedor toma tudo impera e o cidadão apenas acede a alguns conteúdos, num desfecho desigual. É portanto importante perceber os mecanismos da formação de popularidade. Por um lado, a compreensão destes mecanismos ajudará os consumidores a decidir qual o critério de escolha da sua atenção. Por outro, fornece aos divulgadores, empresas de marketing e meios de comunicação, instrumentos de desenho das suas campanhas fundadas na previsão do seu impacto. Por outro ainda, estes estudos podem beneficiar o desenho das novas redes de comunicação, no sentido de as adaptar à procura seletiva de conteúdos e à melhoria do desempenho das *caches* Famaey et al. [2013] [4].

Os estudos dos mecanismos de reputação e de popularidade, também designada por notoriedade, é central na área do Marketing e da Publicidade. Estes estudos são geralmente focados nos mecanismos psicológicos de lembrança das marcas [5], onde os fatores afetivos e emotivos desempenham um papel importante. Outros estudos baseiam-se na quantidade de vendas e em fatores concorrenciais,

[1] http://about.twitter.com/company acedido em Outubro 2014.

[2] http://www.youtube.com/yt/press/statistics.html acedido em Outubro 2014.

[3] http://newsroom.fb.com/company-info/ acedido em Outubro 2014

[4] Ver por exemplo o estudo da empresa MLAB http://www.measurementlab.net (acedido em Outubro 2014) que reporta como os fornecedores de Internet degradam o serviço aos clientes em função da procura dos conteúdos mais populares e pagos.

[5] No Marketing distingue-se a*Notoriedade Top-of-mind* - a primeira referência a uma marca num dado sector, a *Notoriedade Espontânea* - as marcas que para lá da primeira são evocadas espontâneamente e a *Notoriedade Assistida* - quando o reconhecimento das marcas é assistido pelo entrevistador.

como o preço. Estes estudos visam melhorar a publicitação dos produtos e a sua performance no mercado. No entanto, como iremos defender nesta tese, no ambiente de comunicação e para lá de fatores de cativação da atenção, as condições pelas quais a informação flui são determinantes na formação da popularidade. A presente tese procurará demonstrar que as condições em que se processam os fluxos de informação na sociedade contém em si as condições suficientes para a génese da popularidade. A necessidade, involuntária ou não, de cada ator social autónomo mimetizar a palavra de outro e com isso replicar alguns conteúdos e não replicar outros, provoca só por si popularidade, para além das qualidades particulares do que é mencionado [1].

1.2 Principal objetivo da tese

O principal objetivo da tese é a identificação dos processos de formação da popularidade, numa altura em que os meios de produção de conteúdos informativos se democratizaram. Onde antes a produção de informação e publicidade estava centrada num pequeno conjunto de atores na sociedade, nos seus lideres e na comunicação social, assiste-se hoje a uma dinâmica acelerada dos conteúdos e das opiniões e a uma democratização do protagonismo. Neste contexto, é importante perceber como se forma esse protagonismo, quer das pessoas quer dos produtos. Admitimos como ponto de partida que, independentemente das qualidades dos sujeitos, há fatores associados à disseminação da informação que são decisivos na formação popularidade. A pergunta que nos colocámos foi a seguinte:

- *Que mecanismos de propagação da informação têm lugar na formação da popularidade?*

Para abordarmos este problema colocámos uma hipótese e elaborámos vários modelos, que confrontámos com dados experimentais. Estudámos a popularidade de um ponto de vista estático, onde é comparada a popularidade de diferentes

[1] Uma experiência simples consiste em simular um modelo de comunicação em que N agentes, cada um possuindo inicialmente uma mensagem diferente, replicam iterativamente uma mensagem de qualquer um outro agente escolhido ao acaso. Inevitavelmente a partir de certo número de iterações toda a comunidade transmite apenas uma mensagem. Em apêndice apresentamos um modelo em Netlogo onde este fenómeno pode ser observado.

entidades num determinado instante temporal. Estudámos também a popularidade na perspetiva da sua evolução temporal. Finalmente efetuámos um estudo de caso que nos permitiu aprofundar conhecimentos sobre a popularidade e os processos de formação de opinião.

1.3 Contribuições científicas da tese

Tal como na intrincada malha neuronal, que Dennet pretendeu estudar e que constitui o cérebro, a complexidade da malha social onde se produz e consome informação beneficia com a abordagem particular das Ciências da Complexidade. Neste trabalho procurámos aplicar metodologias próprias das Ciências da Complexidade, como as simulações multi-agente, mas também analisar grandes quantidades de dados empíricos para validação dos modelos propostos. Este trabalho de recolha e análise envolveu fases exigentes de qualificação e estruturação de dados, até à construção de software específico, executado durante um período relativamente longo, para recolha dos dados das fontes na Internet (como aconteceu com o tratamento dos dados da Wikipedia e da rede Twitter descritos nas partes III e IV).

As principais contribuições científicas desta tese são as seguintes:

- Um levantamento do estado da arte no âmbito dos estudos sobre popularidade e conteúdos da Internet, nomeadamente no que concerne a previsões sobre o voto e nos estudos sobre influência e dinâmica de opiniões.

- Um modelo explicativo da diferenciação de popularidade entre entidades, independente das suas qualidades próprias e baseado exclusivamente em parâmetros de propagação de informação. Trata-se de um modelo que contempla não só como a popularidade é formada a partir da magnitude anterior da popularidade de cada entidade, como também a proporção de atenção que é prestada entre as entidades já populares e as novas entidades que vão ser objeto de atenção. Este modelo é validado com dados significativos em quantidade, atuais, facilmente verificáveis e acessíveis através da Internet.

- Dois modelos explicativos da evolução temporal da popularidade de uma entidade, concordantes com o modelo anterior na equação elementar de crescimento de popularidade, baseados em parâmetros de difusão de informação. O primeiro modelo contempla um termo que tem em conta a topologia das relações sociais e o impacto de fatores externos que influenciem a evolução temporal da popularidade. O segundo modelo é relativo a um processo endógeno à comunidade, baseado numa multiplicação epidémica com uma taxa de propagação dinâmica. Ambos os modelos são validados com sucesso em dados da evolução temporal de popularidade obtidos de duas fontes: a propagação de *memes* numa rede muito vasta de blogues e a propagação de *hashtags* num conjunto de milhões de *tweets*.

- A confirmação que a popularidade das entidades, candidatos ou partidos, medida pelo número de referências na comunicação social e na rede social Twitter, está intimamente associada à expressão do voto nas eleições. Esta confirmação é obtida através de dados experimentais recolhidos em duas eleições distintas.

- A confirmação que a popularidade destas mesmas entidades está muito correlacionada com a sua referência nos meios de comunicação social. Esta confirmação é efetuada por correlação simples mas também através de simulações de dois modelos multi-agente que confirmam os resultados obtidos experimentalmente.

1.4 Estrutura da tese

A tese está dividida em cinco partes.

Nesta primeira parte é feita a introdução a todo o trabalho de investigação relatado. É definido o âmbito da investigação e os seus principais objetivos. Depois são relatadas as principais contribuições científicas da tese e a sua estrutura.

Na segunda parte é efetuado um levantamento do estado da arte no âmbito da investigação. No capítulo dois examinamos dois conceitos chaves para o trabalho efetuado, necessários para o seu enquadramento, o conceito de Informação e o conceito de Comunicação. No capítulo é feito um levantamento do trabalho

científico sobre o problema da popularidade, procurando examinar as questões que ficaram em aberto e a relação da nossa perspectiva no contexto geral e atual nesta área. Abordamos os conteúdos da Internet como objeto particular de popularidade e fazemos incidir um foco particular no problema da popularidade em política e a relação dos meios *online* com a política.

A terceira parte concentra o núcleo principal desta tese. O quarto capítulo apresenta uma solução para o problema da distribuição da popularidade entre diversas entidades que mostram padrões regulares de distribuição da popularidade no longo prazo. Esta solução tem a forma de um modelos explicativo de mecanismos geradores de popularidade. No quinto capítulo fazemos a validação deste modelo. No capítulo sexto propomos dois modelos adicionais, baseados no processo gerador de popularidade anteriormente descrito, que explicam de modos distintos alguns padrões de popularidade detetados nos dados experimentais. O primeiro modelo é baseado em processos de ramificação e o segundo em processos epidémicos. Seguidamente validamos estes modelos.

Na quarta parte apresentamos um estudo de caso em que examinamos a influência que a grande massa de informação distribuída na sociedade através da comunicação social pode ter na formação da popularidade, nomeadamente na intenção de voto em eleições. No capítulo oitavo detalhamos esse estudo, baseado não só em análise estatística, como em dois modelos de simulação multi-agente que nos permitem abordar e validar experimentalmente níveis de popularidade em varias entidades (candidatos ou partidos) em simultâneo. O estudo centra-se em dois atos eleitorais decorridos no ano 2011. Estes dados experimentais servem de suporte à interpretação de fatores importantes no ato de comunicação constitutivo de popularidade e a sua dependência do fluxo de informação propagado na comunidade pela comunicação social.

Finalmente dedicamos a última parte a discutir os resultados obtidos e a perspetivar a evolução, não só deste trabalho, como do âmbito do estudo e as possíveis aplicações práticas que poderá ter.

Parte II

O Estado da Arte

Capítulo 2

Conceitos básicos

Este capítulo examina com detalhe os conceitos de Informação e de Comunicação ligados ao problema a que se dedica a tese. Este exame permitirá enquadrar o âmbito da análise do problema no estudo geral da propagação da informação, ou seja, encarar os mecanismos de formação de popularidade não em termos das qualidades aparentes dos objectos populares, como é efectuado nos estudos de marketing, mas na forma como a informação sobre esses objectos flui na sociedade através de actos de comunicação.

2.1 Informação

O termo *Informação* designa um conceito abrangente que é predominantemente associado, no uso coloquial, a quantidade de dados, de código ou de texto que pode ser guardada, enviada, recebida ou manipulada, através de diversos meios e suportes. Apesar do papel essencial que desempenha para os seres vivos, quer para os individuos quer para as sociedades, o seu estudo é recente.

Historicamente o estudo da informação pode ser entendido como um esforço para tornar mensuráveis as propriedades extensivas do conhecimento.

Até à segunda metade do século XX quase nenhum filósofo moderno considerou a *informação* como um conceito filosófico importante Adriaans [2012]. Autores antigos como Cícero (106–43 BC) e Santo Agostinho (354–430 AD) discutiram os conceitos platónicos usando os termos *informare* e *informatio* para

traduzir conceitos gregos técnicos, como *eidos* (essência), *idea* (ideia) *typos* (tipo), *morphe* (forma) e *prolepsis* (representação). Os antigos atribuíam à *forma* uma importância relevante. Das quatro causas necessárias para o entendimento dos objectos, Aristóteles (384–322 BC) distinguiu a *Causa Formal*, que consistia na fórmulas e nas suas partes, como a razão 2:1 presente na oitava musical - que caracterizava os objectos. Exemplificando o acto de *informare*, Aristóteles relatou a acção pela qual um sinete de bronze transmitia a sua forma a um pedaço mole de cera, numa analogia com a mente, onde as ideias aí eram impressas e se modificavam umas às outras. A tensão na filosofia clássica entre o idealismo Platónico (*universalia ante res*) e o realismo Aristotélico (*universalia in rebus*) retornou na idade média, relativamente ao problema dos conceitos universais e à questão da existência destes independentemente da matéria. Santo Agostinho descreve, por exemplo, ao estilo de Aristóteles e em analogia à Santíssima Trindade, o processo de apreensão pela visão de uma forma corporal no mundo externo, através da capacidade de *informatio* da visão, da qual resulta a forma existente na mente.

Após a idade média e depois do abandono da problemática da forma e da matéria e dos morfismos entre uma e outra, o conceito de informação ligado à vida mental foi sofisticado. O racionalismo de Descartes (1596–1650) defendeu a noção de ideia inata e apriori, a ideia e a forma como entidades sem tempo. Contrariamente, os empiristas como Locke (1632— 1704) e Hume (1711 — 1776) aceitaram a existência de processos de criação e estruturação das ideias a partir dos sentidos. Com eles a informação poderia nascer na mente, com vários níveis de crença. O dicionário histórico de francês de Godefroy (1881) define por outro lado informação como *action de former*, *instruction*, *enquête*, *science* ou *talent*. Na filosofia moderna o termo foi raramente usado e não foi associado aos processos epistemológicos. A *Informação* passou gradualmente a deter um estatuto de qualidade substantiva, distinta do sentido de processo mental, mas também encarada como uma qualidade do sujeito que é informado. Foi deste modo que seria recuperada pelos cientistas no século XX (Fisher 1925 e Shannon 1948), que inventaram métodos formais para a sua mensuração. O sucesso destes métodos e da sua aplicação prática nas modernas tecnologias da comunicação e da informação levou à emergência de um ramo distinto da filosofia, destinado a analisar o conceito nas suas diversas facetas. Nos dias de hoje a *Informação*

constitui uma categoria central, quer nas ciências quer nas humanidades.

Distinguem-se neste conceito três significados relevantes Adriaans [2012]:

- *Informação* como processo de ser informado - é o seu sentido mais antigo e está associado ao processo de reconhecer uma forma, por exemplo.
- *Informação* como estado de um agente - é o resultado do processo de ser informado.
- *Informação* como a disposição que algo possui para informar - é uma qualidade dos objectos, que pode ser guardada e transportada e que nos permite dizer que algo contém informação.

Depois da II Grande Guerra do século XX, dois trabalhos pioneiros de Claude Shannon Shannon [1948] Shannon and Weaver [1949] deram um significativo avanço ao estudo da informação, através da Teoria Matemática da Comunicação (TMC). Esta teoria é baseada em dois princípios:

- A *Informação* é extensiva - relacionada com a intuição natural de, por exemplo, um texto mais longo conter mais informação.
- A *Informação* reduz a incerteza - relacionada com a intuição de que quando detemos todo o conhecimento possível sobre uma dada circunstância, não podemos obter mais informação sobre esta.

A teoria de Shannon procurou dar resposta a duas questões importantes no domínio das telecomunicações : "Até que ponto pode uma mensagem ser comprimida sem isso afectar o seu conteúdo informativo?"e "Qual a rapidez com que pode ser transmitida uma mensagem num determinado canal sem nenhum erro?"As respostas a estas questões resultam nas definições de entropia H (Equação 2.1) da mensagem e na definição de capacidade do canal C (Equação 2.3) Floridi [2011].

A entropia de uma mensagem M constituida por símbolos x é dada por:

$$H(M) = -\sum_{x \in M} p(x) \log_2 p(x) \tag{2.1}$$

onde a quantidade de informação, ou auto-informação, veiculada por cada símbolo é igual a:

$$I(x) = -\log_2 p(x) \tag{2.2}$$

e é determinada pela probabilidade $p(x)$ do símbolo indeterminado x "acontecer"na mensagem.

No caso de um canal com ruído gaussiano aditivo com uma largura de banda B e uma relação entre a potência do sinal e do ruido gaussiano S/N, a Capacidade do Canal C é dada por:

$$C = B \log_2(1 + S/N) \tag{2.3}$$

Depois de Shannon, outras teorias quantitativas da informação surgiriam. A mais relevante, sendo citada como mais fundamental do que a teoria de Shannon Cover and Thomas [2006], é designada por Teoria Algorítmica da Computação de Solomonof, Kolmogorov e Chaitin Solomonoff [1960], Kolmogorov [1965], Chaitin [1977]. Esta teoria, fundada nos princípios da máquina de Turing e na ciência da computação, aborda a mensagem não do ponto de vista da surpresa e probabilidade de observação, mas naperspectiva da complexidade do programa necessário para a produzir. Quanto mais simples e menos informativa certa mensagem for, menos extenso será o programa necessário para a produzir, e deste modo também menor a sua quantidade de informação em bits. (Este mesmo método de economia é usado frequentemente em ciência, desde Willian de Occam e a sua "Navalha", até ao conceito de Mínimo Comprimento Descritivo de Rissanen Rissanen [1978].)

Outra forma de interpretar o conceito de *Informação* deriva da teoria dos jogos e consiste na estruturação da perceção do mundo, no sentido em que este determina e condiciona a ação. Deste modo, admitindo que o mundo se pode desenrolar num conjunto finito de estados, cada estado constitui uma descrição completa da realidade. Apenas um estado é válido em cada instante. O espaço de estados é um conjunto $\Omega(w)$ que contém elementos $\{w_1, w_2, \ldots, w_n\}$. O lançamento de uma dado, por exemplo, dado todo o restante mundo igual, pode ser representado pelo seguinte conjunto de estados:

$$\Omega = \{1, 2, 3, 4, 5, 6\} \tag{2.4}$$

Uma *estrutura de informação* é o resultado da separação do espaço de estado em sub-conjuntos, a que se chamam *partições* do espaço de estados Birchler and Butler [2007]. No exemplo acima podem ser definidas duas partições "par"e "ímpar", ambas definidas pelas estrutura informativas $\{2, 4, 6\}$ e $\{1, 3, 5\}$, que determinam um certo conhecimento sobre o lançamento dos dados. As partições também definem *eventos* no mundo, resultado do lançamento do dados, que em teoria dos jogos são designados por *conjuntos de informação*. Dizer que um agente detém mais informação que outro, equivale, nesta noção de informação, a afirmar que possui partições mais finas do espaço de estados e que portanto tem mais poder para o descrever em pormenor:

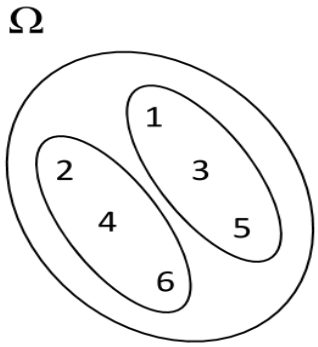


Figura 2.1: Exemplo de uma estrutura da informação "par/impar"aplicada ao conjunto de estados do mundo depois do lançamento de um dado.

Weaver afirmou que "a palavra informação não tem tanto a ver com o aquilo que se diz mas sim com o que se pode eventualmente dizer. A Teoria Matemática da Comunicação lida com os veiculos da informação, simbolos e sinais, e não com a própria informação. Ou seja, informação é a medida da liberdade de escolha quando se seleciona uma mensagem"Weaver [1949]. De fato a Teoria Matemática da Comunicação (TMC) é também designada por uma teoria "sintática", pois a componente da informação associada ao sentido não é por ela enquadrada. A TMC não está interessada no "acerca de", na relevância,na utilidade ou na interpretação da informação. Estes componentes são tratados por outras teorias "semânticas". Apesar da filosofia da informação semântica se ter tornado autónoma da TMC, duas conexões importantes se mantiveram estáveis até à

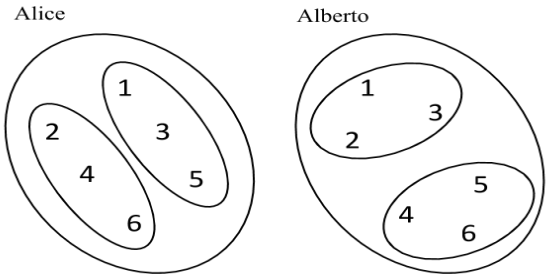


Figura 2.2: Informação heterogénea implica que as estruturas de informação de diferentes agentes se sobrepõem. Alice sabe quando um lançamento de um dado dá "par"ou "impar"enquanto que Alberto sabe quando o lançamento é "alto"ou "baixo".

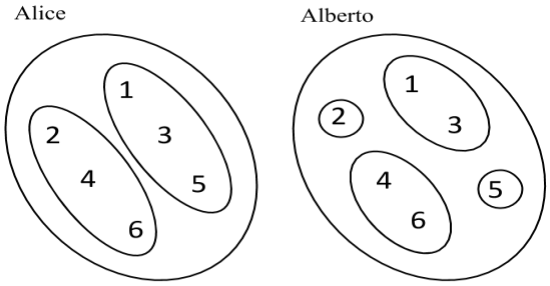


Figura 2.3: Informação assimétrica implica que um agente, aqui Alberto, possui uma estrutura de informação mais fina, ou seja informação superior do que outro, neste caso Alice.

atualidade:

- O modelo de comunicação emissor-canal-recetor
- O chamado Principio da Relação Inversa Barwise [1997] que relaciona a informação com o inverso da probabilidade.

Todas as teorias semânticas da informação incluem estes pressupostos. Admite-se que num extremo do espectro as teorias de informação semântica-factual são fortemente constrangidas pela TMC, noutro extremo elas são constrangidas de uma forma muito fraca Floridi [2011].

Devido à evolução tecnológica, a informação usada entre os humanos tem progressivamente vindo a desmaterializar-se e com isso a ser autonomizada dos meios de transporte comuns. No século XIX ninguém imaginaria ouvir música sem ser perto de um músico. No século passado não imaginaríamos ouvir música sem algum objeto de suporte onde esta estaria alojada - um disco de vinil, uma cassete ou um CD. Com os últimos desenvolvimentos tecnológicos é razoável admitir que podemos ouvir qualquer música, do vasto reportório da criação musical humana, em qualquer parte do mundo onde tenhamos acesso à Internet, a partir de um pequeno telemóvel do tamanho da palma da mão.

Na figura 1 podemos encontrar a evolução cronológica do uso da palavra *informação* e a sua comparação com o uso de outros termos relacionados com transportes, em língua inglesa. É interessante verificar como a variação positiva do uso da palavra, após o início do século XX, é oposta à variação negativa dos termos normalmente associados com o transporte de pessoas e de bens, nomeadamente a variação relativa ao uso do *cavalo* e do transporte ferroviário. Se, por um lado, com a evolução dos transportes o mundo parece mais pequeno, com a evolução das telecomunicações a *informação* parece desempenhar um papel progressivamente mais relevante na vida em sociedade.

Com a massificação das tecnologias da informação a partir dos anos 70, vivemos hoje numa *sociedade da informação*, que alguns autores também designam por *sociedade em rede* Castells [1996], onde o relevo da informação é associado à *comunicação* em rede.

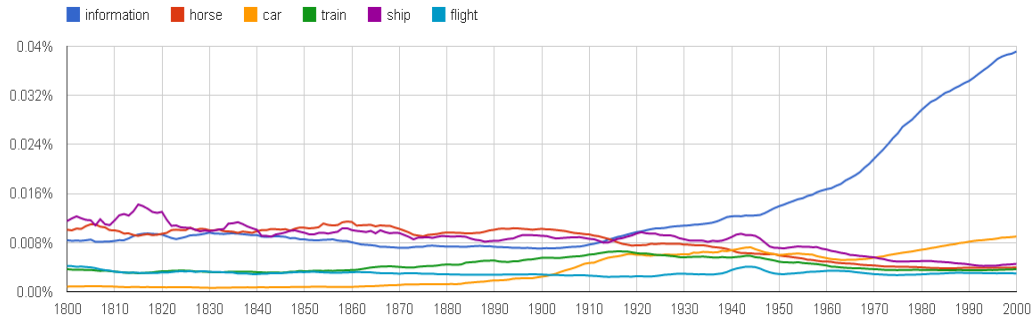


Figura 2.4: Uso do termo *information* versus outros termos relacionados com transportes segundo o Google Ngram Viewer, que reporta a contagem da frequência dos termos num corpus de 4,541,627 livros em Inglês publicados entre as data referenciadas no eixo das abcissas.

2.2 Comunicação

Num artigo clássico sobre teorias da comunicação Craig [1999], Robert Craig sugere sete diferentes perspetivas sobre as quais a comunicação pode ser abordada. Destas perspetivas identificamos a mais relevante para o enquadramento desta tese:

Comunicação no sentido *cibernético* - Comunicação diz respeito ao processamento (transformação, troca, propagação, difusão) de informação que está exclusivamente associada à expressão, à interação e à influência entre pessoas. Neste sentido os problemas relevantes da comunicação são teorizados tendo em atenção as condições em que se dão as trocas de informação, o ruído, a sobrecarga ou sub-carga, a forma e as condições de funcionamento do canal de comunicação.

Palavras-chave associadas a este conceito são por exemplo: *fonte*, *recetor*, *sinal*, *ruído*, *feedback*, *redundância*, *rede*, *função*, mas também *perceção*, *cognição*, *atitude* e *interação*. Os modelos que proporemos mais adiante nesta tese beneficiam desta simplificação metodológica na formulação do conceito de comunicação. No entanto este conceito é multifacetado, multidisciplinar e difícil de sintetizar. Sendo um processo essencial no funcionamento da sociedade e da própria natureza, o campo dos estudos sobre a comunicação, de tão vasto, é considerado por alguns autores como não existente. Tal deve-se ao facto de não haver um

cânone, e consequentemente não haver geralmente contraditórios nem disputa, apesar da vasta literatura publicada Craig [1999]. Contornando desde já o facto de o âmbito da pesquisa sobre comunicação também dizer respeito ao funcionamento da natureza, a nível exclusivamente social e humano, esta área abrange factos tão díspares como o ordenamento jurídico do sistema televisivo; intrincadas operações financeiras em torno de alguns meios de comunicação; episódios de proibição de programas; crises com quedas e triunfos de algumas estruturas produtivas cinematográficas; polémicas sazonais sobre o efeito pernicioso dos *media*, entre outros.

Os *mass media* constituem, simultaneamente, um importante sector industrial, um universo simbólico objeto de um consumo maciço, um investimento tecnológico em continua expansão, uma experiência quotidiana, um sistema de intervenção cultural e de agregação social e uma maneira de passar o tempo Wolf [1985]. Até ao final dos anos 70 os estudos da comunicação foram evoluindo ao ritmo a que esta chegava às massas. Nessa altura todos constataram haver uma insatisfação sobre a falta de sistematização e de um corpo teórico sólido, havendo uma dispersão dos esforços sobre temas variados e o reconhecimento da complexidade dos processos comunicativos. Se até então se distinguiam duas abordagens distintas dos dois lados do Atlântico - uma pesquisa "administrativa"na América, acentuadamente empírica e caracterizada por objetivos cognoscitivos, e uma pesquisa europeia chamada "crítica", teoricamente orientada e mais atenta ás relações gerais existentes entre o sistema social e os meios de comunicação de massa, os termos dessa oposição foram reorientados nos anos 80. A partir dessa altura verificou-se uma confluência entre o interesse na influência dos meios de comunicação sobre o público, tipicamente americana, e a determinação estrutural do pensamento, tipicamente europeia. De igual modo o foco de estudo passou da mera análise da propaganda e da publicidade, para o papel dos meios de informação e da constatação que os modernos meios de comunicação são parte de um único sistema comunicativo, cada vez mais integrado e complexo, requerendo uma abordagem multidisciplinar. Outra das alterações de foco resultou no interesse dos efeitos a longo prazo da comunicação e não meramente nas relações causais imediatas.

Numa fase inicial considerava-se que os processos comunicativos seriam as-

simétricos, envolvendo um sujeito ativo que emitiria um estímulo e um outro passivo que reagiria, que aconteceriam a um nível individual, intencional e episódico. Passou a entender-se o processo comunicativo como efeito a longo prazo, em que as comunicações não intervêm em diretamente no comportamento explícito mas tendem, isso sim, a influenciar o modo como o destinatário organiza a imagem do seu ambienteWolf [1985]. Uma consequência deste entendimento é a passagem do estudo de casos individuais para uma cobertura global, a passagem do estudo com entrevistas para metodologias mais complexas e em vez de serem avaliadas as mudanças de atitude e opinião, procura-se reproduzir o processo em que o indivíduo altera a sua própria representação da realidade social.

Os *mass media*, como um todo, adquiriram um estatuto de construtores da realidade. Neste âmbito uma das hipóteses que ocupa um lugar de destaque é a hipótese do *agenda-setting*, segundo a qual em função dos meios de informação o público passa a saber ou ignorar, presta atenção ou descura, realça ou negligencia elementos específicos dos cenários públicos. Originalmente esta hipótese não defende que os meios de comunicação tentam persuadir o público, mas que são esses meios que determinam a parte da realidade social que lhe é fornecida, sobre a qual ele deve opinar e discutir. Um método de testar a hipótese é verificar se todos os temas abordados pelos *media* são recebidos pelo público. Alguns estudos constatam que existe uma persuasão temperada pela persistência e que nem todos os temas tratados têm o mesmo impacto, ou que podem alterar a ordem de preferências subjetiva do público. A hipótese é aliás mais complexa do que há primeira vista se apresenta. Pode-se perguntar: que conhecimentos e que público para o efeito *agenda-setting*? Por outro lado, terão todos os canais e temas o mesmo efeito? Que processos sub-jazem ao *agenda-setting*? Um assunto é muito importante se for muito mencionado ou é necessário fazer uma análise de conteúdo, de memória e de discurso? A cronologia interessa? A persistência, a ordem temporal; a duração e os intervalos influenciam? Outras questões dizem respeito à caracterização da agenda. Podemos dizer que existe uma agenda intrapessoal - que corresponde ás preferências de cada individuo, uma agenda interpessoal - que corresponde aos temas que ela fala e discute com outros, e uma agenda percecionada - que corresponde à perceção que o sujeito tem da opinião pública. De igual modo outra tripartição diz respeito ao modelo do efeito da

agenda, um modelo de conhecimento - referido à presença ou ausência de um tema na agenda pública, um modelo do realce - que corresponde há existência de dois ou três temas importantes na agenda do público e um modelo das prioridades - que corresponde à hierarquia estabelecida pelos indivíduos num conjunto completo de temas.

Se por um lado a hipótese do *agenda-setting* é bastante sedutora pela sua simplicidade, as variáveis em jogo não o são. O papel do público no modelo, ao invés de passivo e linear, pode ser suposto construtivo num processo coletivo com reciprocidade. Deste modo podemos supor um modelo que se desenrola por fases, uma fase de *focalização* na qual os *mass media* dão relevo a determinada entidade ou personalidade; uma fase de *enquadramento* em que o objeto focalizado é inserido num quadro interpretativo que é intensamente coberto e que é imposto ao público; uma fase de *ligação* em que a entidade ou acontecimento é ligado a um sistema simbólico, de forma que o objeto passa a fazer parte de um panorama social e político reconhecido; e uma fase de *potenciação* em que os protagonistas podem ter a habilidade de influenciar os *mass media* e propor um novo ciclo.

Este carácter reciproco sublinha a importância dos emissores e dos processos produtivos de comunicação de massa nos estudos modernos, e constitui na atualidade outra grande área de investigação. Um dos papeis centrais desempenhados neste âmbito é o de *gatekeeper* (selecionador) que determina os conteúdos que são difundidos e que filtra a agenda. A questão relevante é, então, a maneira como funciona este selecionador - se age conscientemente ou é inconscientemente influenciado, se motivado por razões políticas; jornalísticas ou de negócio, e quais são os critérios na escolha da informação que difunde: *substantivos* - devido aos conteúdos; *materiais* - relativos à disponibilidade e ao meio de comunicação; *dos públicos* - relativos ao público que vai atingir e *concorrenciais* - de razões comerciais e de imagem.

No contexto dos estudos desta tese é importante perceber qual o papel desempenhado pela comunicação de massas na formação da popularidade. No capítulo 5 vamos examinar a influência da comunicação na formação de popularidade através do estudo de um caso de eleições em Portugal.

2.3 Conclusão

As duas secções anteriores examinaram dois conceitos necessários para fundamentar o estudo da popularidade. Se por um lado a informação constitui a matéria viva que permite a comunicação, por outro sem comunicação não faz sentido dizer que algo é popular.

Do ponto de vista informativo, partiremos das propriedades da informação para definir um conceito de mensagem que permita a formação de popularidade. Do ponto de vista da comunicação, analisaremos o processo através do qual a troca de mensagens permite essa formação.

No capítulo seguinte abordaremos a investigação já efetuada sobre popularidade e enquadraremos a presente tese nesse trabalho.

Capítulo 3

O estudo da popularidade

Após abordarmos dois conceitos fundamentais necessários para a análise dos fenómenos associados à difusão de informação na sociedade, vamos examinar e reportar neste capítulo o trabalho, que no nosso melhor conhecimento, tem sido efetuado com a finalidade de explicar e de prever a popularidade.

Uma destas tarefas a que se tem dedicado a investigação é a de encontrar padrões e regularidades nos dados observados na sociedade. Muitos fenómenos sociais e humanos têm sido descritos através de leis de potência. Desde da regra de 80-20 de Vilfredo Pareto (1896); da lei de Zipf na frequência relativa das palavras em textos Zipf [1949] ou da contagem de nomes distintos em fabricantes e distribuidores Zipf [1950]; de citações de artigos em jornais académicos Simon [1955]; dos rendimentos, capacidades industriais e tamanhos de firmas Simon and Bonini [1958] Stanley et al. [1995] Ijiri and Simon [1977] Okuyama et al. [1999]; da dimensão das populações nas cidades Gabaix [1999]; da colaboração entre actores Barabási and Albert [1999]; dos dividendos em corridas de cavalos Park and Domany [2001]; ou das distribuições dos rendimentos individuais Persky [1992], as distribuições da classe das leis de potência são comuns no estudo dos sistemas complexos Newman [2005]. Todos estes fenómenos apresentam regularidades a um nível "macro", mas envolvem num nível "micro"uma miríade de pequenas escolhas e interações feitas por indivíduos diversos, com expectativas e interações díspares. No entanto, as regularidades existem e mostram-se empiricamente evidentes.

Em Bass [1995] Frank Bass define a generalização empírica como "um padrão

ou uma regularidade que se repete em diferentes circunstâncias e que pode ser descrito matematicamente, de modo gráfico ou por métodos simbólicos". O modelo de Bass Bass [1969] da difusão de inovações, ou o modelo Dirichlet de consumo de marcas Goodhardt et al. [1984], constituem dois exemplos de generalizações empíricas no âmbito do consumo. É no âmbito da Economia que se desenvolveram os primeiros trabalhos de investigação sobre a preponderância das marcas e dos produtos.

3.1 Popularidade, marketing e a publicidade

Uma das abordagens iniciais na procura de regularidades iniciou-se com o trabalho de Robert Gibrat de 1931 *Inégalités Économiques* Gibrat [1931]. Gibrat apresentou a Lei dos Efeitos Proporcionais, que pretendia explicar a dimensão relativa das firmas e da estrutura industrial. Sendo inicialmente um modelo exclusivamente económico, veio a encontrar eco em muitas disciplinas como modelo de crescimento. Herbert Simon e Charles Bonnini Simon and Bonini [1958] afirmavam em 1958 que os muitos estudos até então efetuados sobre o crescimento de firmas eram "monotónicamente similares", chegando invariavelmente ao resultado que identifica poucas firmas grandes e uma maioria de mais pequenas. Os anos 50 e 60 seriam profícuos em novos modelos Sutton [1997], nomeadamente estocásticos, compilados por Simon e os seus co-autores Ijiri and Simon [1977] em 1977. Nos anos 80 desenvolveu-se uma nova literatura em reação a modelos puramente estocásticos e na procura de melhorar a qualidade dos dados econométricos Sutton [1997]. Apesar de não dizerem diretamente respeito ao tema da popularidade, estes modelos de crescimento precedem o modelo que iremos apresentar nesta tese.

Os estudos de mercado e a popularidade estiveram sempre ligados. A principal atividade do *marketeer* é auscultar o mercado que pretende satisfeito, fidelizado e alargado. O publicitário, através do seus meios, procura que o mercado percecione o valor dos produtos ou serviços e em consequência aumente a sua popularidade. A gestão da marca, o *branding*, iniciada em 1931 por um publicitário da multinacional Procter e Gamble, Neil McElroy, ao advertir os seus colegas que as marcas deveriam ser trabalhadas independentemente do produtos, define com a maior

pertinência a importância da popularidade na sociedade moderna.

De entre os estudos existentes sobre generalizações empíricas e procura de leis na área das marcas e produtos, destaca-se um relatório de 1976 no Boston Consulting Group Stern and Deimler [2006] efetuado por Bruce Henderson que estipula a existência sistemática de 3 marcas fortes em qualquer sector, um líder e dois competidores. Outro dos estudos é de 1981 Buzzell [1981] onde Robert Buzzell avança com um modelo no qual a popularidade das marcas, revelada pela fatia de mercado, obedece a uma lei logarítmica similar à da dimensão das firmas. Em 1984 Goodhardt et al. [1984] usam a distribuição de Dirichlet para modelar as regularidades nas quotas de mercado. Em 2004 Kohli e Sah Kohli and Sah [2004] mostram que a distribuição através de centenas de mercados de marcas obedece a uma lei de potências. Este resultado deriva da aplicação de um modelo similar ao que propomos, publicado em 1974 por Bruce Hill Hill [1974], que atribui uma distribuição Dirichlet-multinomial à probabilidade dos atos de compra serem atribuídos a produtos, da qual resulta uma lei de potência no valor das suas vendas. O modelo que será aqui proposto, baseado na lei de Gibrat, acomoda melhor os dados experimentais nos casos de popularidade que estudamos a uma distribuição lognormal. No capítulo 6 discutiremos as particularidades das duas abordagens.

Por outro lado, com o advento da era da comunicação à escala global, para além da comunicação entre as empresas e os indivíduos, a troca de informação entre estes no seio da sociedade mostrou ser extremamente relevante em certos fenómenos de génese de popularidade. Um exemplo paradigmático deste caso é o fenómeno dos vídeos virais, como o vídeo *Gangnam Style*[1] na plataforma Youtube, que contabilizava em Abril de 2014 1.9 biliões de visionamentos provocados apenas por citações Jiang et al. [2014]. Outro exemplo são os *memes* gerados na Internet, que consistem em atividades, conceitos ou expressões propalados pela Internet, como aconteceu com a dança do *Harlem Shake* [2] em 2013. Este processo de criação de popularidade através da palavra trocada entre indivíduos é aproveitado pelos *marketeers* como técnica publicitária Leskovec et al. [2007a].

[1] http://en.wikipedia.org/wiki/Gangnam_Style acedido em Outubro 2014.

[2] http://en.wikipedia.org/wiki/Harlem_Shake_(meme) acedido em Outubro 2014.

A presente tese incide sobretudo na popularidade de conteúdos divulgados na Internet. Estes conteúdos constituem ativos importantes no mundo digital em que vivemos e são um recurso económico cada vez mais significativo. Neste âmbito, interessa-nos a relação entre o utente da Internet e o conteúdo através da medida da sua presença. Esta medida pode ser feita através da contagem do número de comentários, do número de partilhas ou de ligações. Optámos por utilizar o número de visionamentos, visitas ou menções de entidades no Twitter.

3.2 Popularidade, redes sociais e Internet

Os primeiros estudos sobre popularidade de conteúdos na web reportam-se aos anos 90, quando foram realizados os primeiros cálculos do acesso, não a conteúdos mas a páginas inteiras, notando-se uma distribuição relativa com bastante obliquidade aproximada pela lei de Zipf Tatar et al. [2014]. Estes trabalhos procuraram melhorar o tráfego na Internet através da implementação das primeiras estratégias de *catching*. Seguidamente, alguns investigadores focaram-se nos conteúdos. Com o surgimento das plataformas da Web 2.0, ricas em dados sobre os utilizadores, foram abordados os problemas das conexões entre utilizadores e da difusão do mesmo conteúdo em diferentes domínios.

Do ponto de vista dos objetos de popularidade na Internet podem-se distinguir:

- Os vídeos - onde se destaca a plataforma Youtube. Os diversos estudos nesta plataforma não são unânimes acerca da distribuição da popularidade. Assim, dependendo do conjunto de dados usados, obtiveram-se distribuições em lei de potência com um corte superior exponencial Cha et al. [2009], Weibull Cheng et al. [2008], Gama Cheng et al. [2007] ou Lognormal Borghol et al. [2011] (como nos nossos estudos, ver Parte III). A evolução temporal da popularidade dos vídeos é muito diversa. Crane e Sornette Crane and Sornette [2008] descobriram uma evolução temporal em três modos similar a alguns dos padrões exibidos na presente tese.

- Notícias - As notícias dos media são uma das principais fontes de informação na Internet e são largamente disseminadas nas redes sociais. A sua popula-

ridade é difícil de medir, sendo para isso por vezes usados os comentários. As distribuições de popularidade correlacionadas variam também conforme os trabalhos: lei de potências Mishne and Glance [2006a]Tatar et al. [2012] ou lognormal Tsagkias et al. [2010]. Em termos de evolução temporal a popularidade das notícias dura regra geral menos tempo que a dos videos.

- *Bookmarks* - os *sites* de *bookmarking* Digg, Slashdot ou Reddit, apesar da sua pouca divulgação em Portugal, têm uma grande divulgação nos EUA. Nestas plataformas os utilizadores interagem através de comentários e ligações a conteúdos, que discutem e classificam. A evolução temporal da popularidade é mais rápida do que a das notícias, sendo os conteúdos populares em média de 6 horas e desvanecendo-se num dia Szabo and Huberman [2010]. As distribuições encontradas para a popularidade são Weibull Wang et al. [2012] e lognormal Szabo and Huberman [2010]Lerman and Ghosh [2010] Van Mieghem et al. [2011].

- *Posts* - As redes sociais permitem a interação a partir de pequenos conteúdos partilhados entre grupos utilizadores. As redes mais populares são o Facebook, o Twitter e a rede chinesa Weibo com mais de 300 milhões de utilizadores ativos. A estrutura destas redes segue uma lógica de amizade e de difusão limitada a círculos de indivíduos. Cada utilizador pode colocar conteúdo na rede, que depois pode ser replicado. Alguns estudos sobre o Twitter reportam que os conteúdos são efémeros, replicados nos instantes seguintes a serem colocados e raramente repetidos no dia seguinte Yang and Counts [2010]Zaman et al. [2013]. A maior parte do estudos atribui uma distribuição em lei de potência à popularidade do *tweets* Ma et al. [2013a]Hong et al. [2011]Kupavskii et al. [2013]

.

A área de investigação sobre redes sociais incide sobre temas como os seguintes:

- Popularidade
- Propagação e difusão de informação
- Influência

- Comunidades
- Previsão
- Sistemas de recomendação
- Análise automática de sentimentos

O domínio dos padrões de popularidade e divulgação dos conteúdos trazem benefícios para diversos atores: os utilizadores podem filtrar melhor a informação, os produtores e os fornecedores podem organizar e entregar melhor os seus conteúdos e os publicitários podem melhorar as suas campanhas. Em termos gerais, no âmbito da popularidade, a investigação procura dois propósitos: encontrar métodos e algoritmos para efetuar previsões e descobrir perfis e padrões. Muitos artigos partem das propriedades da difusão de informação e da relação entre utilizadores para construir modelos explicativos dos padrões encontrados. Outros procuram tratar os dados agregados com algoritmos de *Data Mining* ou outras técnicas de inferência estatística para a determinação de processos de previsão.

Na área da política os estudos sobre padrões vão desde campanhas no Facebook Williams and Gulati [2007] Williams and Gulati [2008] e no Twitter Shamma et al. [2009], à relação sincrónica destes dados com os debates televisivos Diakopoulos and Shamma [2010]. Alguns estudos usam análise automática de sentimentos para efectuar previsões de voto O'Connor et al. [2010]Choy et al. [2011]. Esta técnica também é utilizada para avaliar o comportamento dos mercados Bollen et al. [2011a], das marcas Jansen et al. [2009b] ou dos eventos noticiosos Thelwall et al. [2011]. A popularidades dos eventos e das notícias, dada a utilização intensiva das redes online, permite correlacionar os dois fluxos. Exemplo disso são os estudos de audiências a partir do uso das redes de Twitter Shamma et al. [2010] Lin et al. [2014] Gupta et al. [2012] Bandari et al. [2012] Matsubara et al. [2012] ou da plataforma Digg Lin et al. [2013] Szabo and Huberman [2010]. Uma das áreas de fértil investigação é a previsão de temas populares, os chamados *trending topics* Hong et al. [2011]Petrovic et al. [2011]Kong et al. [2012] Gao et al. [2014] ou o uso de palavras chave Ma et al. [2012]Lin et al. [2013]Ma et al. [2013b]Kong et al. [2014a]Kong et al. [2014b]. Mas nem só a plataforma Twitter é utilizada

como objecto para a investigação e recolha de dados experimentais. As redes de citação de artigos ciêntíficos Brody et al. [2006]; a popularidade de temas colocados em blogues Gruhl et al. [2004]Yang and Leskovec [2011]Kim et al. [2011]; dos temas e notícias mais frequentes em jornais Tatar et al. [2011]Castillo et al. [2014]; dos artigos da Wikipedia Ratkiewicz et al. [2010]; do visionamento na rede Youtube Pinto et al. [2013]Ahmed et al. [2013], a popularidade dos utilizadores rede MySpace Couronne et al. [2013] ou a popularidade de conteúdos na web em geral Lee et al. [2010] são outros exemplos de objecto de investigação na área dos media e das redes sociais *online*.

Em resumo, as plataformas sociais *online*, como novo ambiente de comunicação na sociedade, têm constituído nos anos mais recentes um terreno fértil de investigação na área da comunicação e da relevância social dos temas, dos produtos e das pessoas.

3.3 A popularidade e o voto

Estudos recentes, baseados na análise de sentimentos e emoções, introduziram novas técnicas na deteção de tendências e na previsão de efeitos a partir da análise dos conteúdos e das ligações entre utilizadores nas redes sociais. Neste caso procura-se não só descrever o comportamento social ou observar o impacto de uma determinada campanha publicitária, mas também prever o futuro.

Dada a natureza da fonte de informação, estas previsões: Yu and Kak [2012]:

- Dizem respeito a eventos relacionados com atividade humana. Eventos climatológicos ou acidentes naturais, pela sua natureza, não são previsíveis por este meio, apesar de recentemente epidemias terem sido alvo de investigação Achrekar et al. [2011]Santos and Matos [2013].

- São baseados em amostras representativas do mundo real, pressupondo-se que dados recolhidas em grande quantidade constituem amostras significativas.

- Dizem respeito a eventos facilmente verbalizados em público. Eventos pouco

discutidos em sociedade dificilmente são analisados através da conversa social.

As previsões através das redes sociais efetuaram-se em diversos campos.

No *Marketing* descobriram-se evidências fortes da correlação entre picos na escala de vendas entre produtos e o número de *posts* em blogues relacionados com esses produtos.Gruhl et al. [2005]. Os consumidores com opiniões extremas acerca de produtos tendem a expressa-las nas redes sociais. Por outro lado, os consumidores com opiniões moderadas normalmente não as expressam. As organizações e as marcas são referidas em cerca de 19% de todos os *tweets*, de entre os quais 80% não expressam nenhum sentimento vincado Jansen et al. [2009a]. No entanto o fenómenos do passa-palavra é muito influente na compra de um produto ou serviço, especialmente se for efetuado a partir de amigos ou de pares Domingos and Richardson [2001]Engel et al. [1969]Katz and Lazarsfeld [1970]. Sabe-se também que a opinião negativa tende a ter mais influência do que a positiva Skowronski and Carlston [1989]Duan et al. [2008]Park and Lee [2009]. Não só o consumo mas também a adoção de novos produtos é afetada pelas relações sociais em redes. Vários estudos comprovam que a escolha individual de um produto é influenciada pelos amigos Granovetter [1978]Bhatt et al. [2010]Backstrom et al. [2006]. Particularmente sabe-se que as pessoas com menos amigos são mais influenciáveis Bhatt et al. [2010]. Todos estes factos levam a crer que a predição das condições de mercado para certas marcas pode de fato ser potenciada através da análise dos media sociais. Por este motivo, vulgarizaram-se as plataformas de *media analytics* que permitem aos *marketeers* um acompanhamento de perto da presença e da notoriedade das suas marcas.

Uma situação semelhante à do Marketing acontece na previsão da *venda de bilhetes de cinema*. A previsão, não só das vendas, como também da classificação dos filmes e do número de ecrãs, é um tema bastante estudado Sharda and Delen [2006]Zhang and Skiena [2009]Asur and Huberman [2010]Simonoff and Sparrow [2000]Mishne and Glance [2006b] e tem tido bastante sucesso. Este sucesso deve-se ao facto dos filmes, sendo marcas, serem amplamente discutidos na esfera pública - estima-se que cada filme chegue a ser comentado em mais de 100.000 *tweets* nos Estados Unidos Asur and Huberman [2010]. Também os da-

dos de audiência, classificações e receitas são amplamente estudados e divulgados, permitindo aos investigadores a obtenção de resultados relevantes.

A outro nível, as previsões no âmbito *Macroeconomia*, quer a nível regional, quer a nível nacional ou global, têm sido experimentadas com menos sucesso. Os investigadores têm-se debruçado sobre índices económicos O'Connor et al. [2010] através da análise de sentimentos, mas com resultados pouco estáveis no tempo. O caso mais estudado é o do mercado de capitais onde, apesar da hipótese da eficiência apontar para a impossível previsão, alguns resultados chegaram a ser obtidos. Um exemplo foi o estudado no quadro de mensagens do portal financeiro da *Yahoo* Antweiler and Frank [2004] com resultados estatisticamente identificáveis mas relativamente insignificantes no contexto da magnitude habitual da volatilidade deste domínio. Um aumento de 100% no conteúdo relacionado nos media sociais equivaleu a apenas 0.2% de diminuição do preço. No entanto o mesmo estudo revelou que o volume de tráfego nos media é preditivo da volatilidade do mercado. Também o emprego de certas palavras com carga emotiva se correlaciona melhor com o índice bolsista do que o intervalo de tempo de reação a uma mensagem, ou a importância de cada emissor medida pelo número de relações de amizade Zhang et al. [2011]. O sentimento expresso tem um valor preditivo. Apesar da generalidade dos sentimentos, positivos versus negativos, não demonstrarem especial correlação com a evolução do mercado, dois estudos mostraram que sentimentos específicos, como calma, alegria ou ansiedade estavam associados às cotações de fecho do índice Dow Jones e do S&P 500 Bollen et al. [2011b]Gilbert and Karahalios [2010]. Um estudo mais geral, focado nos termos de pesquisa no motor Google, mostrou uma correlação significativa entre as cotações dos mercados e por exemplo a palavra 'dívida' Preis et al. [2013].

Outra área de investigação na previsão através de redes sociais é a da disseminação de informação. O entendimentos dos processos de disseminação de informação é importante pois permite antecipar e compreender fenómenos como os boatos e o impacto da publicidade e das notícias. Em termos genéricos, todo o comportamento social tem por suporte a comunicação e a difusão de informação. Esta difusão pode ser analisada a um nível micro, entre pares, ou a um nível macro, em grandes cadeias chamadas *cascatas* Leskovec et al. [2007b], onde o mesmo conteúdo pode ser espalhado por dezenas ou centenas de milhares de pes-

soas. Utilizando as características das pessoas na rede social, é possível mapear certo tipo de difusão de informação em função destas características Zaman et al. [2010]. Na propagação em grande escala, em redes sociais como a *Digg*, o *Twitter* ou o *YouTube*, sabe-se que os tempos de vida da informação são diferentes Szabo and Huberman [2010]Lerman and Galstyan [2008] Lerman and Hogg [2010]. A previsão com base no conteúdo é mais fácil quando os tópicos têm tempo de vida curto, como acontece na *Digg*, pois este tempo não é contaminado com o efeito da propagação na rede. A contrário, no *YouTube* um vídeo pode continuar a ser visitado por vários meses ou anos. De igual modo, a popularidade dos conteúdos pode ser correlacionada com outros parâmetros, como a hora ou o dia da semana, mas de um modo geral a análise em função do conteúdo é mais fiável quando a propagação é rápida e ainda não filtrada pela rede de utilizadores.

Finalmente, o comportamento eleitoral também tem sido objeto de previsões a partir de redes sociais. Neste âmbito efetuámos um estudo de caso, a partir de uma recolha de conteúdos publicados na rede *Twitter* nos meses que antecederam as eleições presidenciais e legislativas de 2011, que nos permitirá situar melhor a questão central da presente tese. Este estudo de caso será apresentado na Parte III do documento de tese.

Quando em 2008 a campanha nas redes sociais do presidente Obama ultrapassou as expetativas em termos de donativos e de mobilização de massas, a possibilidade da comunicação online de influenciar os resultados eleitorais alertou o público em geral para o potencial das novas plataformas de comunicação. A rede *Twitter* passou a ser considerada um meio de debate na arena pública. No entanto, esta relação entre tecnologias e política é mais antiga. Antes do resto do mundo, em 1996, os candidatos presidenciais americanos possuíam já páginas na Internet para promoção das suas campanhas. Em 1998 o candidato do partido reformista Jesse Ventura utilizou o correio eletrónico para vencer as eleições estaduais no Minnesota. Em 2000 John McCain efetuou grande parte da angariação para o financiamento da sua campanha presidencial através da Internet. A partir de 2004 até aos dias de hoje, blogues por todo o mundo discutem 24 horas por dia o mundo político Tumasjan et al. [2010].

Toda esta atividade política é realizada com o propósito da troca e da publicitação, quer do ponto de vista pessoal quer institucional, de opiniões e visões.

Com a aceleração e a popularidade das novas tecnologia, multidões de indivíduos rapidamente são contactados pelo processo de comunicação, multiplicando o poder coletivo das mensagens Shirky [2008]. Na semana de 17 de Janeiro de 2001, durante o julgamento de impedimento do presidente filipino Joseph Estrada, 7 milhões de mensagens de texto SMS levaram à rua milhares de filipinos contra o seu presidente corrupto. Desde então, diversos movimentos públicos de rua foram despoletados por comunicações eletrónicas no mundo inteiro: Espanha, 2204; Bielorrússia, 2006; Moldávia, 2009; Irão 2009, Tailândia 2010; a Primavera Árabe em 2011, em que os media sociais se transformaram em ferramentas de coordenação para movimentos políticos, enquanto líderes políticos autoritários os tentam controlar Shirky [2011]. O mais significativo sinal deste esforço de controlo é o projeto chinês Escudo Dourado, chamado em jargão a Grande *Firewall* da China, que envolve milhares de trabalhadores e enormes recursos computacionais para censurar a Internet na China.

A expressão de opiniões nos media sociais pode ser vista como uma forma racional de participação política, onde os benefícios dessa participação excedem em grande medida os custos associados. A comunicação política através das redes sociais tem menos custos pois é efetuada em ambientes relativamente amigáveis, ao contrário do que seria em confrontos face a face Goidel [2011]. Algumas pesquisas revelam que os utilizadores de estatuto social mais elevado utilizam a Internet com um propósito de reforço do seu capital social. Este facto acentua a fratura informacional entre estratos sociais e que deve ser considerada na análise deste media. De igual modo, estudos recentes revelaram diferentes modos de utilização para diferentes níveis de empenho político. Segundo estes estudos, os utilizadores mais empenhados expressam-se mais frequentemente e com mais veemência Kirzinger [2011]. Este fato constitui um problema de desvio na sondagem de opinião,que tem paralelo nas sondagens por telemóvel devido à sua utilização ser preferida em idades mais jovens e com menos recursos. Como nas sondagens clássicas, o problema pode no entanto ser atenuado através da atribuição de fatores de ponderação. Ao contrário do que acontece nas sondagens clássicas, onde a percentagem de respostas ou o número de respostas completas pode não ser suficiente, na análise de media sociais esse problema não existe. No entanto o fato de não se poderem formular questões articuladas pode impedir certo tipo

de análises. Por outro lado, a interrogação através do que já foi publicado não padece do problema da resposta ser uma "questão pessoal", que muitas vezes condiciona os resultados em sondagens clássicas Greenberg [2011]. Os proponentes da opinião deliberativa clássica defendem que uma opinião pública relevante só emerge após um processo deliberativo em que as visões opostas são confrontadas e portanto a discussão pública nos media sociais tende a ser mais realista do que a sondagem pessoal. [1]

A magnitude da amostra de um determinado conteúdo político pode refletir o seu sucesso eleitoral. Nas eleições presidenciais primárias nos Estados Unidos em 2008, o número de apoiantes na rede *Facebook* refletiu o resultado final Williams and Gulati [2008]. Nas eleições federais alemãs de 2009 um método similar foi utilizado com sucesso na rede *Twitter*, apesar de apenas 4% dos utilizadores serem responsáveis por mais de 40% do conteúdo Tumasjan et al. [2010]. O sentimento das mensagens pode também auxiliar as previsões, apesar de não ser significativo O'Connor et al. [2010]. No entanto existem argumentos contra estes processos de predição. Em eleições provinciais na Columbia Britânica em 2001 procurou-se prever o resultado com base em posts num forum web, mas com maus resultados Jansen and Koop [2005]. Em alguns trabalhos, ao contrário daquele que efetuámos, os dados recolhidos não são datados e não é possível compará-los com resultados de sondagens clássicas. Do mesmo modo, numa das pesquisas foram excluídos partidos mais pequenos Jungherr et al. [2012].

Deve-se notar que os media sociais não refletem a demografia da sociedade. Em termos de idade os votantes nos Estados Unidos no ano 2000 repartiam-se em 36% entre os 18 e os 24 anos, 50% entre os 25 e os 34, e 68% tinham mais de 35 anos Metaxas et al. [2011] mas no *Twitter* mais de 60% dos utilizadores tem menos de 24 anos Cheng et al. [2009]. No entanto, o mesmo não acontece com a localização e em outras redes este inconveniente não existe. No caso do

[1] Alguns académicos defendem o oposto - a expansão da discussão pública a redes sociais tende a atrair mais a visões particulares do que a incrementar o diálogo e a discussão. Isto deve-se ao fato da leitura de perspetivas em competição tender a aumentar o entendimento das mesmas mas diminuir o engajamento político. Assim um eleitorado mais polarizado e empenhado não é necessariamente mais informado, deliberativo ou reflexivo. A desvantagem é que a polarização tende a incrementar a participação, especialmente a um nível mais básico, e este processo pode distorcer a realidade reportada nos media sociais Mutz [2006]. A experiência reportada no nosso estudo na comunidade portuguesa de *Twitter* é semelhante (ver Parte III).

Facebook todos os principais dados pessoais dos utilizadores são recolhidos, assim como outra informação relativa a modos de vidas, que a própria empresa utiliza para a segmentação da sua atividade de marketing. O maior problema em termos de amostragem é a cobertura do universo real dos utilizadores.

Sabe-se que mera escolha do incumbente nas eleições pode indicar com relevância o vencedor. Por exemplo, nas eleições para o congresso americano em 2008, 91,6% dos candidatos incumbentes ganharam. Essa percentagem foi de 84,5% em 2010. A indicação do incumbente torna possível obter uma precisão na previsão superior a 80%. No entanto, em comparação com a precisão obtida na rede *Twitter*, a precisão com este método é pior com um erro médio de 17% quando se contabiliza apenas o volume de menções, de 7,6% com análise de sentimentos e de 2-3% nas sondagens clássicas Metaxas et al. [2011] Gayo-Avello et al. [2011]. Apesar de todos estes inconvenientes e de não haver nenhum estudo que indique um intervalo temporal necessário na qual basear a recolha de dados da rede social, baseamos o nosso estudo numa recolha mínima de três meses, que comparámos cronologicamente com o resultados das sondagens clássicas e em que obtivemos resultados relevantes (ver Parte II da tese).

3.4 Conclusão

Dada a sua utilidade na economia de mercado, na forma indireta do sucesso de produtos, das pessoas e das marcas, a popularidade tem sido objeto de investigação no seio dos mercados de consumo. Recentemente, com a divulgação dos novos meio de comunicação e a facilidade de recolha de grandes quantidades de dados, também os conteúdos mediáticos têem sido estudados. Tanto quanto nos é possível conhecer, os resultados obtidos centram-se na definição de padrões e de algoritmos preditivos que procuram, com mais ou menos sentido prático, dominar o fenómeno da difusão da informação em diversos meios. Todos estes estudos se restringem a contextos definidos e não procuram a generalização. Apesar dos estudos de Marketing e Publicidade, e alguns no âmbitos da Internet, abrangerem com mais pertinência a importância dos conteúdos e dos fatores psicológicos e sociológicos que determinam a recetividade e a divulgação, não é do nosso conhecimento um estudo mais geral que procure analisar o papel exclusivo da

transmissão da informação na génese da popularidade.

A presente tese procura abordar o problema de um ponto de vista geral. Pretendemos definir o processo da forma mais abstrata que for possível. Deste modo definiremos um veiculo geral do processo de formação da popularidade - a *mensagem*, e em função desta são definidos dois modelos para os dois eixos ortogonais de evolução temporal: a quantificação da popularidade entre entidades num dado instante e a dinâmica da popularidade para uma só entidade.

Esta tese centra-se num processo dinâmico que procura explicar como a popularidade se origina num meio coletivo através da propagação da informação de forma multiplicativa, resultando a distribuição dos seus efeitos numa distribuição probabilística lognormal. Os modelos aqui propostos procuram facilitar a elaboração teórica e promover consequências testáveis para lá dos dados experimentais usados.

Parte III

Modelação de Popularidade

Capítulo 4

Modelo de Distribuição da Popularidade

De acordo com o estado da arte apresentado nos capítulos anteriores, a investigação sobre as distribuições de popularidade não é unânime. Se por um lado predominam as distribuições em lei de potência, os ajustamentos que são feitos aos dados experimentais são frequentemente apenas visuais, sem uma quantificação precisa da melhor função que se ajusta aos dados [1]. Por outro lado, estas distribuições não são justificadas pela discrição do mecanismo que as originam. Para responder à hipótese central desta tese, procurámos investigar esse mecanismo, tentando posteriormente identificar o seu ajuste aos dados experimentais que recolhemos. Com este objectivo, elaborámos e propomos um modelo de propagação suportado numa abstração do ato comunicativo e na mensagem trocada por esse ato. Partimos do pressuposto de que a popularidade é construida a partir da informação trocada na comunidade, cujo veículo é por essência a mensagem. O modelo é construído de um modo progressivo, sendo elaborado na medida da validação das suas várias fases. Começamos por uma equação simples a partir da qual chegamos à formula final.

[1] Ver por exemplo o artigo de Clauset et al Clauset et al. [2009] que propõe um método mais preciso para este ajuste e que utilizamos na validação dos modelos propostos nesta tese

4.1 Descrição do Modelo de Distribuição de Popularidade

A definição do Modelo de Distribuição de Popularidade parte do conceito de mensagem. A definição é baseada na abordagem sobre o conceito de informação já exposto. Propomos na secção seguinte uma definição de mensagem adequada à interpretação dos dados experimentais, à investigação sobre popularidade e ao modelo de distribuição de popularidade introduzido na tese.

4.1.1 Definição de Mensagem

- *Uma mensagem é uma unidade de comunicação trocada num acto comunicativo.*[1]

Reportando-nos ao esquema clássico da comunicação, emissor/canal/receptor e à breve abordagem à informação do capitulo 2, definimos deste modo *mensagem* no sentido de contentor para a informação que é trocada na comunicação e também como unidade indivisivel. A relação entre a mensagem e a informação que ela transporta não é no entanto direta.

No sentido semântico da informação, que é aquele que nos interessa aqui explorar, podem-se distinguir dois tipos de informação: a *informação procedimental* e a *informação factual* Floridi [2011]. A informação procedimental diz respeito a instruções para a acção, a informação factual diz respeito à informação sobre factos ou crenças. Os dois tipos de informação são relevantes para efeito do estudo dos mecanismos de popularidade. Factos e ações são igualmente importantes para a construção da popularidade. Pensemos, por exemplo, no que acontece com as multidões quando as palavras de ordem levam a uma actuação concertada.

Uma mensagem pode ser mais ou menos informativa - com maior ou menor número de bits. A quantidade de informação da mensagem, dependente do grau de surpresa para o receptor, pode variar com o contexto. Um único bit de informação "sim/não"que responde à questão "O desembarque dos aliados está eminente?"pode ser codificada com longas cadeias de simbolos, como o foi quando

[1]Definição presente na Wikipedia (http://en.wikipedia.org/wiki/Message) acedida em Junho 2014

a BBC difundiu uma mensagem com duas linhas de um poema em 1 Junho de 1944. As duas linhas tinham várias dezenas de bits Floridi [2011]. De igual modo a mensagem associada a um único clique/bit numa determinada página de um browser de internet pode provocar a descarga de diversos mega bits de informação contida num livro digitalizado.

Podemos considerar que, para efeitos da propagação da informação associada a popularidade, as mensagens que a veiculam são codificadas de tal modo que essa codificação não interfere na génese da popularidade. De outro modo teríamos de analisar cada mensagem por si. Assim, o objectivo é o de propôr uma formula genérica de mensagem.

Na informação veiculada pelas mensagens considera-se que os dados nelas contidos se apresentam ao receptor como possibilidades restritoras ("constraining affordances") que permitem gerar conhecimento a partir da informação. Os dados, considerados como possibilidade restritoras - respostas aguardando as questões relevantes - são transformados em informação através de um processamento semântico (a questão relevante é associada à pergunta certa) num certo *nível de abstração*(NdA). Estes níveis de abstração constituiem interfaces que medeiam a relação entre o observador e aquilo que é observado Floridi [2011]. Como vimos no capitulo 2 a respeito das estruturas de informação e das partições do espaço de estado, os níveis de abstração dependem do receptor e não são necessáriamente hierárquicos, mas são entre si comparáveis [1]. A informação assim obtida, uma vez justificada e verdadeira, transforma-se em conhecimento do agente.

O nosso objectivo é o de analisar as trocas de informação de um ponto de vista genérico, independentemente das caracteristicas particulares do emissor ou do receptor. Para atingir este objectivo é necessário ignorar os níveis de abstração a partir dos quais os dados são transformados em informação. Deste modo, dizemos que a definição de mensagem que nos interessa contém apenas dados que são interpretados genéricamente por qualquer receptor. De igual modo existe um debate em aberto sobre o papel do valor de verdade no conceito de

[1]Por exemplo a informação que um agente recebe quando ouve no meio de uma frase o nome de um pintor famoso é diferente entre se o recetor for critico de arte experimentado ou um leigo em pintura.

informação. Apenas a informação verdadeira, que se pode verificar, conta como informação factual? Neste caso as tautologias, necessáriamente verdadeiras ou as contradições, não seriam informativas. Podemos no entanto observar que para a análise da popularidade pode interessar informação cujo valor de veracidade não é conhecido. Recordemos para isso a função dos slogans publicitários que muitas vezes se reduzem apenas a uma palavra.

Assim podemos melhorar a nossa definição de mensagem do seguinte modo:

- *Definição de Mensagem - Uma mensagem, para efeitos da análise dos mecanismos da popularidade, é uma unidade de comunicação que é trocada num acto comunicativo. É constituida exclusivamente por dados, codificados de forma a serem passíveis de uma interpretação genérica. A informação que veiculam não é necessáriamente verdadeira e pode ser procedimental.*[1]

A função da mensagem na génese da popularidade é essencial. A popularidade da entidade referida pela mensagem advém da popularidade, ou seja do número de receptores, que captam a mensagem. Deste modo ela pode ser aproximadamente medida pelo número de receções, independentemente do nível de abstração em que é interpretada por cada um.

4.1.2 A génese da popularidade

Uma constatação contraintuitiva que poderemos desde já sublinhar é a seguinte:

- *É condição necessária e suficiente para haver popularidade que o conjunto de mensagens trocadas numa comunidade fechada seja finito.*

De facto, se todas as mensagens tiverem exactamente a mesma probabilidade de repetição, ou seja, tiverem o mesmo valor intrínseco para cada individuo, e em qualquer circunstância, mas se dois individuos repetirem por acaso em simultaneo a mesma mensagem, a probabilidade dessa mesma mensagem ser posteriormente repetida é maior.

[1] Exemplos de mensagens são o nome de uma pessoa, a imagem de uma pessoa, de um local ou qualquer outra entidade nominal, um slogan publicitário, um logotipo, uma cor, um gesto, uma linha melódica ou uma história, mas não um facto. Neste sentido o conceito de mensagem aproxima-se do conceito de *significante* no modelo linguístico de Saussure.

Este processo evolui ao ponto de, dado tudo o resto igual, tender a apagar da comunidade todas as mensagens com possibilidade de repetição inferior, as que inicialmente apenas foram repetidas por um individuo, até remanescer única na sociedade [1].

A realidade é diferente, pois cada mensagem tem uma possibilidade mais acentuada do que outra de ser repetida. Esta possibilidade depende do sentido subjectivo da mensagem e das circunstâncias da sua repetição e origina uma distribuição não trivial de popularidade de mensagens.

De notar, no entanto, que as condições para existir a diferenciação são triviais e elementares.

4.1.3 Propagação e difusão

A propagação de uma mensagem no seio de uma sociedade pode ser comparada ao progresso de uma epidemia ou da divulgação da adopção de um novo produto Easley and Kleinberg. No inicio do séc passado, em epidemiologia, foram propostos modelos para a propagação de agentes patogénicos numa população Anderson and May [1991]. Destes, destaca-se o modelo SIR [2] que representam a dinâmica dos diversos compartimentos populacionais portadores, suscetíveis ou recuperados da doença. Inicialmente, até aos anos 90 Rothenberg [2007]Valente and Pumpuang [2007], estes modelos não contemplavam a estrutura das conexões entre os atores sociais, pois admitiam que cada ator podesse infetar um qualquer outro aleatóriamente. Com o advento dos estudos de redes sociais (ver capítulo 2) os modelos passaram a contemplar a estrutura do tecido social no que concerne aos padrões de contacto e à natureza das doenças. Uma doença infeciosa, como a gripe, possui um rede de contactos susbtancialmente diferente da rede emergente a partir de um virus sexualmente transmissível, como o HIV. Esta diferença de estruturas no tecido social, onde se dá a propagação, representa um grau adicional de sofisticação e de detalhe dos modelos.

No modelo a seguir apresentado optámos por considerar o caso da difusão generalizada, em que cada agente pode passar a mensagem a qualquer outro. Em

[1] Ver o modelo de simulação multi-agente Netlogo em apêndice

[2] 'Suscepible, Infected, Recovered', assim como os modelos 'Susceptible, Infected, Removed'.

seguida o modelo é adaptado a condições particulares da rede social. Na figura 4.1 encontra-se exemplificado um processo de comunicação de uma determinada mensagem utilizando as entidades *emissor* e *receptor*.

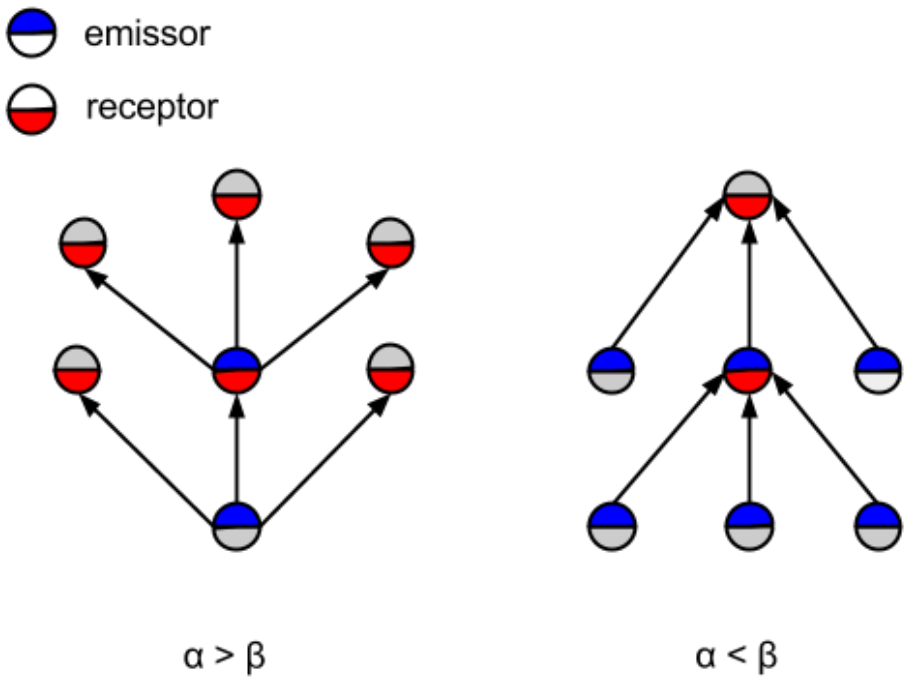


Figura 4.1: Propagação de uma mensagem numa comunidade de agentes e a sua dependência dos parâmetros α_i e β_i.

Para delimitar o âmbito da análise, assumimos que existe um conjunto finito M de mensagens e que a cada mensagem correspondem 3 parâmetros, por sua vez associados, cada um deles, a um agente/ator pertencente a uma comunidade. Estes 3 parâmetros são os seguintes:

- α_{ij} é a probabilidade da mensagem m_i ser escutada pelo agente a_j.
- β_{ij} é a probabilidade da mensagem m_i ser repetida pelo agente a_j.

- θ_{ij} é a probabilidade da mensagem m_i ser esquecida pelo agente a_j.

Deste modo, independentemente do contexto no qual decorre a comunicação, cada mensagem é autónomamente caracterizada pelo percurso que efetua quando é trocada entre os agentes. Conforme atrás referido, uma mensagem contém apenas dados. No entanto, é a informação a estes associada, interpretada por cada agente, que determina a magnitude dos parâmetros.

Admitimos nesta categorização que a probabilidade de uma mensagem ser escutada tem sobretudo a ver com o estímulo que esta apresenta para o individuo. Assim, uma mensagem mergulhada num ruído comunicacional muito forte, ou uma mensagem pouco clara, tem menor probabilidade de ser escutada do que uma mensagem objectiva, atractiva e pertinente.Inevitávelmente, componentes contextuais determinam esta probabilidade. No entanto, supomos um valor médio da probabilidade de uma mensagem ser escutada, num conjunto de indivíduos. Por outro lado a probabilidade de a mensagem ser repetida tem a ver com o valor que a sua difusão pode ter para o emissor. No caso de um orgão de comunicação social é ao editor que cabe esse papel. No caso da publicidade ao director de campanha. No caso do individuo é a ele próprio, em função do valor que a mensagem representa na sua situação particular. Também neste caso os componentes contextuais são importantes, mas admitimos que para cada mensagem existe um número médio de individuos numa amostra significativa de uma população que a repetem se ela for escutada.

Com exclusão de circunstâncias individuais e contextuais, a forma como a mensagem é recebida ou escutada é essencialemente determinada pelo suporte da mensagem:

- Na televisão
- Na rádio
- No computador
- No telefone
- Em papel

- Na rua

Por outro, a probabilidade de a mensagem ser repetida depende sobretudo do alvo do canal através do qual é veiculada. Por ordem descrescente de probabilidade:

- Publicidade, anúncios, campanhas, promoções.
- A comunicação social.
- Eventos públicos, conferências, festas, comícios, rituais, arte.
- Reuniões, redes sociais online.
- Conversas, diálogos, chat online.

Poderíamos dizer, fazendo um paralelo com as causas primeiras de Aristóteles, que a probabilidade da mensagem ser escutada tem uma forte dependência da sua causa material, da sua causa eficiente e da sua causa formal. Depende daquilo de que é feita - o seu suporte, de quem a fez - o seu autor, da sua forma - como se apresenta. A probabilidade de a mensagem ser replicada depende da sua causa final, aquilo para que foi feita - o seu público e as suas consequências.

Em qualquer destes casos, existem, em abstrato, dois modos distintos de propagação da mensagem:

- *De um agente para vários outros*, como é o caso dos orgãos de comunicação social ou num discurso dirigido a um grupo de pessoas.
- *De um agente para outro*, como é o caso de um diálogo.

No entanto, em termos teóricos e para efeitos do estudo da popularidade, a sua diferenciação não é importante, pois uma determinada mensagem chega, na sua máxima difusão, a um sub-conjunto da sociedade. Num caso podemos admitir que chega um grupo com N pessoas, noutro que apenas chega a uma. Estas duas modalidades estão por si englobadas nas definições de α e β.

Não é propósito deste estudo analisar os efeitos da memória. A memória poderia ser importante se abordássemos a construção da Memória Coletiva. No

entanto, o nosso estudo não foca o parâmetro θ, centrando a sua atenção nos outros dois parâmetros.

Resumindo, o conceito de mensagem, conforme foi atrás definimos, permite-nos formalizar os nossos modelos em função do veiculo de informação sobre a entidade à qual queremos analisar a popularidade. Assim, se uma mensagem consistir por exemplo numa fotografia, a popularidade que analisamos é a popularidade das entidades presentes nessa fotografia. Se for um vídeo, a popularidade que analisamos é a de todos os significados, utilizando a terminologia semiótica de Saussure, dos quais o vídeo é significante. Se for um nome de utilizador numa rede social, é essa pessoa. Se for um slogan publicitário, todas as associações a esse slogan, etc. Desta forma, adaptamos a noção geral de popularidade das entidades do mundo real à possibilidade de a informação sobre elas ser difundida e transmitida.

4.2 Formalização do Modelo de Distribuição de Popularidade

No caso da distribuição da popularidade entre entidades, interpretamos de um modo genérico os parâmetros das mensagens atrás referidos, em termos de propagação num modo estático, ou seja analisamos a popularidade das diversas mensagens atingida até um determinado instante temporal. Os agentes são indistintos, não possuindo conexões particulares entre si e admitimos que o parâmetro θ não influencia o modelo. O que nos interessa é ser a popularidade relativa das mensagens entre elas.

Depois de eliminados os factores de variabilidade com os agentes, o modelo fica da seguinte forma:

- α_i é a probabilidade de a mensagem m_i ser escutada.
- β_i é a probabilidade de a mensagem m_i ser repetida.

A melhor forma de entendermos a infuência destes parâmetros na difusão das mensagens é através da análise gráfica do seu impacto. Na Figura 4.1 encontram-

se sumáriamente desenhados os percursos hipotéticos de propagação de uma mensagem.

Se M_i for o evento, a mensagem m_i foi escutada e M_i^* a mensagem m_i foi repetida, então:

$$P(M_i^* \mid M_i) = \frac{P(M_i \mid M_i^*)\beta_i}{\alpha_i} = \beta_i \tag{4.1}$$

Como a mensagem não pode ser repetida sem ser escutada, podemos considerar os dois parâmetros independentes:

$$P(M_i^* \cap M_i) = \alpha_i\beta_i \tag{4.2}$$

O número de agentes efectivamente receptores da réplica de uma mensagem m_i é então proporcional ao produto dos dois parâmetros. Vamos chamar a este produto γ_i e à popularidade da mensagem associada P_i :

$$P_i \stackrel{def}{=} \mid M_i \mid\sim \alpha_i\beta_i \tag{4.3}$$

$$\gamma_i \stackrel{def}{=} \alpha_i\beta_i \tag{4.4}$$

Como as mensagens se vão replicando a partir dos recetores das mesmas, o incremento da popularidade de cada uma depende também da quantidade de receptores já existentes.

Podemos representar este processo na seguinte equação diferencial:

$$\frac{dP_i}{dt} = \gamma_i P_i \tag{4.5}$$

Um modelo formalmente semelhante foi proposto em 1931 pelo engenheiro francês Robert Gibrat, quando propôs uma lei para o crescimento das firmas que se ajustava aos dados conhecidos na época Gibrat [1931]. A lei de Gibrat, que ficou a ser conhecida como 'Lei do Efeito Proporcional', inspirou-se no trabalho de Jacobus Kapteyn. Este investigador estava interessado no aparecimentos de muitas distribuições enviesadas, especialmente em biologia, que atribuía ao efeito aditivo de muitas pequenas influências aleatórias que, operando independente-

mente, geram uma distribuição gaussiana Sutton [1997].

A forma mais simples de apresentar o modelo de Gibrat é notar o tamanho de uma firma num determinado instante t por x_t e atribuir a uma variável ϵ_t a taxa proporcional de crescimento entre o período $(t-1)$ e t, da forma:

$$x_t - x_{t-1} = \epsilon_t x_{t-1} \tag{4.6}$$

então:

$$x_t = (1+\epsilon_t)x_{t-1} = x_0(1+\epsilon_1)(1+\epsilon_2)\dots(1+\epsilon_t) \tag{4.7}$$

se considerarmos intervalos de tempo curtos, é natural admitirmos ϵ_t muito pequenos, justificando a aproximação $\ln(1+\epsilon_t) = \epsilon_t$. Aplicando logaritmos obtemos:

$$\ln x_t = \ln x_0 + \epsilon_1 + \epsilon_2 \cdots + \epsilon_t \tag{4.8}$$

Se admitirmos que os incrementos ϵ_t são variáveis independentes, com uma média m e uma variância σ^2, assimptóticamente, quando $t \to \infty$, o termo $\ln x_0$ tenderá a ser insignificante comparado com $\ln x_t$, de forma que é natural admitir $\ln x_t$ tendo uma distribuição gaussiana com média mt e variância $\sigma^2 t$. Por outras palavras, a distribuição no limite de x_t é lognormal.

A equivalência entre a equação 4.5 e a equação 4.6 é imediata. Ou seja, é de esperar que, obedecendo a popularidade de uma certa mensagem P_i a uma lei de efeito proporcional, a distribuição a longo prazo das popularidades seja lognormal.

Podemos agora sofisticar o modelo, admitindo que a popularidade de cada mensagem P_i é construida a partir de unidades de atenção que lhe são prestadas, conforme vimos atrás na definição de α_i.

Recorrendo a uma evolução do modelo original de Gibrat Growiec et al. [2008], fazemos uma paralelo para o caso de mensagens sobre entidades e consideramos que cada mensagem é sujeita individualmente à atençao repetida de cada pessoa e que esta atenção pode ser variável. Um exemplo disso acontece quando um indivíduo vê um filme uma única vez ou por várias vezes. Podemos assim admitir

dois pressupostos de partida:

- A atenção prestada a cada mensagem é proporcional à quantidade de atenção que já lhe foi prestada, ou seja, a quantidade de unidades individuais de atenção, prestadas por diferentes indivíduos, aumenta proporcionalmente ao número de pessoas que já prestaram atenção à mesma mensagem.
- A magnitude da atenção prestada por cada pessoa varia, obedecendo a uma taxa de variação aleatória.

Formalizando, podemos dizer existem $N(t)$ mensagens m em que, cada uma, num determinado instante t, possui $K_m(t)$ unidades de atenção. No instante $t = 0$ existem $N(0)$ mensagens correspondendo a $n(0)$ unidades de atenção. Em cada intervalo temporal uma nova unidade de atenção é prestada. Deste modo, no instante t existem $n(t) = n(0) + t$ unidades de atenção distribuídas pelas várias mensagens.

Por outro lado, consideramos que com uma dada probabilidade ρ a nova unidate de atenção vai para uma nova mensagem, com uma probabilidade μ uma mensagem deixa de ter atenção e com a probabilidade λ a nova unidade de atenção vai para uma mensagem já existente.

Adicionalmente, supomos que a unidade de atenção atribuída às mensagens existentes segue uma lei de atribuição preferencial às mensagens com mais atenção com a probabilidade $P_m = \lambda K_m(t)/n(t)$.

Consideramos também que cada unidade de atenção tem uma intensidade aleatória x_i, que é independente do número de unidades de atenção de cada mensagem $K_m(t)$. Ou seja, cada mensagem m tem $K_m(t)$ unidades de atenção $x_i(t), i = 1, 2, \ldots K_m(t)$ onde K_m e $x_i > 0$ são variáveis aleatórias independentes. A cada instante de tempo $t + 1$, o tamanho de cada unidade de atenção é aumentado ou diminuido por um factor $\gamma_i(t) > 0$, de forma que:

$$x_i(t) = \gamma_i(t) x_i(t-1) \tag{4.9}$$

Verificamos de novo $\ln x_i(t) = \ln \gamma_i(t) + \ln x_i(t-1) = \sum_{\tau=0}^{t} \ln \gamma_i(\tau)$ e pelo Teorema do Limite Central, $x_i(t)$ tem uma distribuição Lognormal.

Resolvendo o modelo em etapas, consideramos primeiro o caso mais simples em que todas as mensagens possuem o mesmo número de unidades de atenção, ou seja, quando $K_m = K$ não é uma variável aleatória.

- Seja $K = 1$. Neste caso cada mensagem possui apenas uma unidade de atenção, que é a atenção de um único individuo. Como vimos anteriormente, as unidades de atenção obedecem a um crescimento proporcional, ou seja, a equação de distribuição das magnitudes de x_i aproxima-se no limite de uma distribuição lognormal: $\ln x_i(t) \sim N(tm_{\gamma_i}, t\sigma^2_{\gamma_i})$.

- Seja $K > 1$. Neste caso cada mensagem tem um número igual de unidades de atenção pela comunidade. $X(t) = \sum_{i=1}^{K} x_i(t)$ é uma soma de variáveis aleatórias que, conforme o caso anterior, possuem uma distribuição assimptótica lognormal. A soma de variáveis lognormais não tem uma fórmula fechada. Ben Slimane Ben Slimane [2001] apresenta no entanto limites superior e inferior para esta soma:

$$1 - \left[\Phi(\frac{\ln x - m_X}{\sqrt{\sigma^2_X}})\right]^K \leq P(\sum_{j=1}^{K} x_i(t) > x) \leq 1 - \left[\Phi(\frac{\ln(x/K) - m_X}{\sqrt{\sigma^2_X}})\right]^K \quad (4.10)$$

onde Φ denota a função cumulativa da distribuição Normal padronizada $m_X = E(\ln x_i(t)) = tm_{\gamma_i}$ e $\sigma^2_X = Var(\ln x_i(t)) = t\sigma^2_{\gamma_i}$. Ou seja, a função complementar cumulativa de probabilidade situa-se entre a potência K de duas funções complementares cumulativas lognormais.

- Seja $K \to \infty$. Ou seja, cada mensagem tem uma grande exposição a todos os indivíduos. Neste caso, pelo Teorema do Limite Central a distribuição da quantidade de atenção é assimptóticamente Gaussiana:

$$\frac{\sum_{j=1}^{K} x_i(t) - K\mu_x}{\sqrt{K}\sigma^2_x} \to N(0,1) \quad (4.11)$$

à medida que as unidades de atenção que cada mensagem possui crescem para infinito, a distribuição da atenção agregada nas mensagens aproxima-se de

uma distribuição Gaussiana com média $\mu_X = K\mu_x = Ke^{t(m_{\gamma_i}+\sigma^2_{\gamma_i}/2)}$ e variância $\sigma^2_X = K\sigma^2_x = Ke^{2t(m_{\gamma_i}+\sigma^2_{\gamma_i}/2)}(e^{t\sigma^2_{\gamma_i}} - 1)$. Verificamos que μ_X e σ^2_X crescem linearmente com K mas exponencialmente com t, ou seja, a convergência para a lognormal devida ao processo de Gibrat, à aleatoriedade proporcional da atenção, é muito mais rápida do que a convergência para a Gaussiana devido ao aumento do número de unidades de atenção de cada mensagem.

O caso mais normal consiste em admitir que K_m não é fixo e que $\lambda > 0$ e $\mu > 0$, ou seja que cada mensagem tem um valor diferente de unidades de atenção, que pode vir a crescer e que novas mensagens entram no sistema e saiem.

Para obtermos a distribuição a longo prazo da popularidade das mensagens, teremos de calcular:

$$P(X > x) = \sum_{K_m=1}^{\infty} P(K_m) P(\sum_{j=1}^{K_m} x_i(t) > x) \tag{4.12}$$

Seguindo ainda Growiec et al. [2008], admitimos que o último fator é aproximado pela inequação de Slimane 4.10 da seguinte forma:

$$P(\sum_{j=1}^{K_m} x_i(t) > x) = 1 - h(x)^{K_m} \tag{4.13}$$

$$h(x) = \Phi\left(\frac{\ln(x/K_m^l) - m_X}{\sqrt{\sigma^2_X}}\right) \tag{4.14}$$

onde $l \in [0, 1]$ representa um fator de ponderação entre as duas fronteiras da inequação. Não dependendo $P(K)$ de x a função densidade de probabilidade obtém-se pela derivação da equação 4.12:

$$P(x) = h'(x) \sum_{K_m=1}^{\infty} K_m h(x)^{K_m-1} P(K_m) \tag{4.15}$$

Para calcularmos $P(K_m)$, a distribuição do número de unidades de atenção pelas mensagens, temos que examinar a formulação do modelo. Fu et al. Fu et al. [2005] e Yamazaki et al. Yamasaki et al. [2006], no âmbito do crescimento de firmas e não admitindo a probabilidade de extinção, obtiveram uma solução

aproximada para esta distribuição quando o número de mensagens inicial é finito, que no limite $t \to \infty$ corresponde a uma lei de potências com um corte superior exponencial.

$$P(K_m) \approx \frac{1}{\lambda} K_m^{-(1+\frac{1}{\lambda})} \int_0^{K_m} e^{-y} y^{\frac{1}{\lambda}} dy \tag{4.16}$$

Reed e Hughes Reed and Hughes [2004], por outro lado, no âmbito da distribuição do tamanho de genes e admitindo extinção, encontraram uma solução aproximada no caso de $N(0) = 1$ e $n(0) = 1$ que corresponde também a uma lei de potências:

$$P(K_m) \approx \frac{\rho}{\lambda}(1 - \frac{\mu}{\lambda})^{-\frac{\rho}{(\lambda-\mu)}} \Gamma\left(1 + \frac{\rho}{(\lambda - \mu)}\right) K_m^{-\left(1+\frac{\rho}{(\lambda-\mu)}\right)} \tag{4.17}$$

Se admitirmos $K_m \to \infty$ $\mu = 0$ e $\rho = 1$, as duas soluções equivalem-se. Neste caso temos que a equação 4.15 vem:

$$P(x) = h'(x)\frac{\rho}{\lambda}(1 - \frac{\mu}{\lambda})^{-\frac{\rho}{(\lambda-\mu)}} \Gamma\left(1 + \frac{\rho}{(\lambda - \mu)}\right) \sum_{K_m=1}^{\infty} h(x)^{K_m - 1} K_m^{-\frac{\rho}{(\lambda-\mu)}} \tag{4.18}$$

Se considerarmos o caso em que não há extinção:

$$P(x) = h'(x)(\frac{\rho}{\lambda})^2 \Gamma\left(\frac{\rho}{\lambda}\right) \sum_{K_m=1}^{\infty} h(x)^{K_m - 1} K_m^{-\frac{\rho}{\lambda}} \tag{4.19}$$

Aproximando o somatório por um integral:

$$\sum_{K_m=1}^{\infty} h(x)^{K_m - 1} K_m^{-\frac{\rho}{\lambda}} \approx \int_1^{\infty} h(x)^{(s-1)} s^{-\frac{\rho}{\lambda}} ds \tag{4.20}$$

$$\approx \frac{1}{h(x)} \left(\int_0^{\infty} h(x)^s s^{-\frac{\rho}{\lambda}} ds - \int_0^1 h(x)^s s^{-\frac{\rho}{\lambda}} ds \right) \tag{4.21}$$

$$\approx \frac{1}{h(x)} \left(\frac{\Gamma(1 - \frac{\rho}{\lambda}) - \gamma(1 - \frac{\rho}{\lambda}, -\ln h(x))}{-\ln h(x)^{(1-\frac{\rho}{\lambda})}} \right) \tag{4.22}$$

$$\approx \frac{1}{h(x)} \frac{\Gamma(1 - \frac{\rho}{\lambda}, -\ln h(x))}{-\ln h(x)^{(1-\frac{\rho}{\lambda})}} \tag{4.23}$$

Considerando esta aproximação quando $\rho \ll \lambda$, ou seja quando a porção do aparecimento de novas mensagens é muito inferior ao número das existentes :

$$\frac{\Gamma(1-\frac{\rho}{\lambda}, -\ln h(x))}{h(x)} \approx \frac{e^{\ln h(x)}}{h(x)} = 1 \tag{4.24}$$

Reduzimos a equação 4.19 ao seguinte modelo:

$$P(x) \approx (\frac{\rho}{\lambda})^2 \Gamma(\frac{\rho}{\lambda}) \frac{h'(x)}{-\ln h(x)^{(1-\frac{\rho}{\lambda})}} = C(\rho, \lambda) \frac{h'(x)}{-\ln h(x)^{(1-\frac{\rho}{\lambda})}} \tag{4.25}$$

Observamos que o modelo se ajusta bem aos valores experimentais. Este resultado, semelhante ao encontrado por Growiec Growiec et al. [2008] no caso de firmas, excepto no que respeita aos coeficientes de proporcionalidade entre a escolha de novas mensagens ou de existentes, representa um estiramento da função lognormal $h'(x)$ que é escalada inversamente pela sua função cumulativa. A função 4.25 sofre uma alteração da sua forma relativamente à sua curva tal que existe uma sobrevalorização ou sobvalorização da função para valores elevados de x que depende do valor da potência. Na figura 4.2 podemos observar o efeito deste factor multiplicativo para $\lambda = 2\rho$ e para $\lambda = 2/3\rho$. Quanto à probabilidade de novas mensagens terem mais atenção, com ρ maior, em detrimento de velhas mensagens, a probabilidade das mensagens com menor popularidade aumenta significativamente. Ao contrário, quando a probabilidade da atenção se foca nas mensagens existentes, as popularidades mais elevadas são mais prováveis.

Obtemos assim um modelo inicial de popularidade que prevê uma distribuição lognormal estirada pelo conjunto de mensagens que referem entidades, segundo a definição que atrás adiantámos, com um estiramento dependente da proporção de atenção dos agentes.

De seguida validamos este modelo através de um conjunto de dados recolhido na Internet. Para isso vamos recorrer a um *dataset* constituído por o número de visitas a páginas da Wikipedia agrupadas por sectores e a um outro constituído pelo número de visionamentos de dois conjuntos de vídeos na plataforma Youtube. Com ambos os conjuntos de dados procuramos ajustar a curvas da função 4.25 por um processo de minimização dos erros quadráticos.

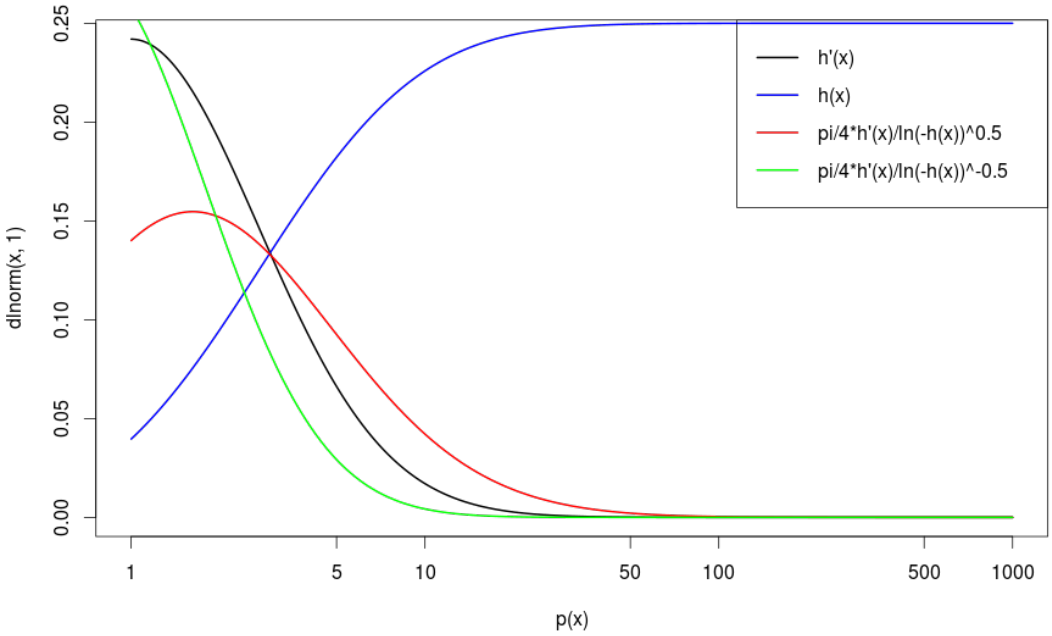


Figura 4.2: Gráfico da função lognormal standard ($\ln N(1.0, 1.0)$) e da função afectada pelo factor multiplicativo especificado na equação 4.25.

Capítulo 5

Validação do Modelo de Distribuição da Popularidade

A validação do modelo proposto no capítulo anterior é efetuada através do ajuste de dados experimentais à função do modelo (equação 4.25). Para cada serie de dados este ajuste é efetuado de duas formas.

O primeiro ajuste é tentado na distribuição complementar acumulada da popularidade dos dados experimentais. Neste caso é tentada a função que melhor se ajusta à curva: se lei de potências, se exponencial ou se lognormal. O método usado é o preconizado por Clauset et al. Clauset et al. [2009]. Este método preconiza minimizar a distância de Kolmogorov-Smirnov para valores crescentes de x_{min}, ou seja para valores crescentes da popularidade, até cada uma das funções minimizar o ajuste. A função cujo x_{min} é menor e cuja estatística KS é menor é a função que se considera melhor adequada aos resultados.

O segundo ajuste é tentado pelo método dos mínimos quadrados, desta vez na função densidade de probabilidade dos dados experimentais. Neste caso é aplicada a equação do modelo uma vez que na fase seguinte se verificou que o melhor ajuste foi sempre, e com bastante distância, à função lognormal. Neste ajuste são então determinados os parâmetros da equação 4.25 que melhor correspondem aos dados.

A validade e qualidade do modelo é corresponde portanto ao ajuste com menores erros quadráticos, ou seja com menor soma quadrática dos resíduos. Os testes que de seguir relatamos são portanto idênticos, no entanto aplicados a diversas

realidades afim de reforçar o nosso argumento.

5.1 Ajustamento do modelo a série de cantores/compositores listados na Wikipedia

O primeiro conjunto de dados de validação é constituído por uma série com o número de visitas a páginas da Wikipedia. Este primeiro conjunto diz respeito a 1963 cantores/compositores americanos listados na Wikipedia, na lista de cantores compositores americanos (ver Anexo **B**. Os dados foram recolhidos no mês de Junho de 2014 e denotam a popularidade dos cantores traduzida no número de visitas a cada página. [1] A distribuição acumulada da sua popularidade está representada na figura 5.1.

Conforme referimos são testados ajustamentos de diferentes funções à curva complementar acumulada da distribuição da popularidade. As estatísticas de Kolmogorov-Smirnov e os respectivos parâmetros para os diferentes ajustamentos estão reportadas na tabela 5.1. Considerando que x representa a popularidade retirada dos dados experimentais, as funções densidade de probabilidade que são testadas são as seguintes:

Distribuição Lognormal com média μ e variância σ^2 :

$$p(x;\mu,\sigma) = \frac{1}{x} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}} \quad , x > 0 \tag{5.1}$$

Distribuição em Lei de Potência com expoente α :

$$p(x;\alpha) = x^{-\alpha} \tag{5.2}$$

Distribuição Exponencial com taxa λ :

$$p(x;\lambda) = e^{-\lambda x} \quad x \geq 0 \tag{5.3}$$

O teste de Kolmogorof-Smirnov Clauset et al. [2009] permite medir o bom

[1] O site que fornece estas estatísticas mantém um top (http://stats.grok.se/en/top acedido em Junho 2014) das páginas da Wikipedia mais visitadas cuja quantidade acompanha frequentemente a escala de popularidade das noticias do dia.

ajustamento (*goodness of fit*) das curvas experimentais ao modelos formais. Na prática este teste mede a máxima distância entre a curva teórica e os valores experimentais, sendo uma das medidas mais usadas para testar o ajustamento. Os ajustamentos testados foram efectuados para um valor mínimo de popularidade. Como podemos observar, a curva que melhor se ajusta e com o valor mínimo mais baixo, portanto que se ajusta a mais pontos, é a curva lognormal.

Distribuição	Estatística KS
Lognormal $Pi_{min} = 2.7$	0.022
Lognormal $Pi_{min} = 100$	0.022
Exponencial $Pi_{min} = 1248.4$	0.036
Lei de Potência $Pi_{min} = 5.4$	0.075

Tabela 5.1: Valores mínimos da distância de Kolmogorov-Smirnov para os valores de Pi_{min} usados.

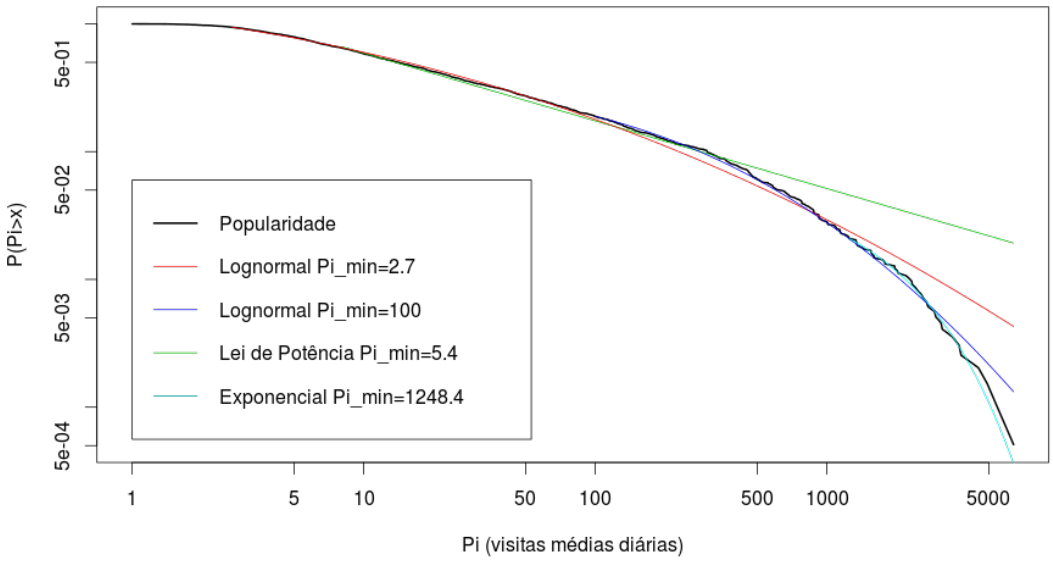


Figura 5.1: Função distribuição complementar acumulada da popularidade das visitas, pela média diária, das páginas de um conjunto de 1963 cantores-compositores americanos. Ajustamentos a funções de distribuição lognormal, de lei de potências e exponencial para o sector da curva superior a $P_i$min

Na figura 5.2 podemos observar o ajustamento à equação que obtivemos com uma soma quadrática de resíduos de 0.0009493 por comparação com 0.00137247 para a distribuição lognormal simples. O erro padrão dos resíduos foi de 0.001548 com 375 graus de liberdade. Nota-se que a cauda possui um estiramento que é melhor acomodado pela curva da equação 4.25 do que pela curva da equação lognormal simples. De facto, tratando-se de uma lista pouco modificada ao contrário do exemplo seguinte, uma vez que poucos novos cantores são introduzidos nesta lista, $\rho \ll \lambda$ e a popularidade elevada tende a ser valorizada.

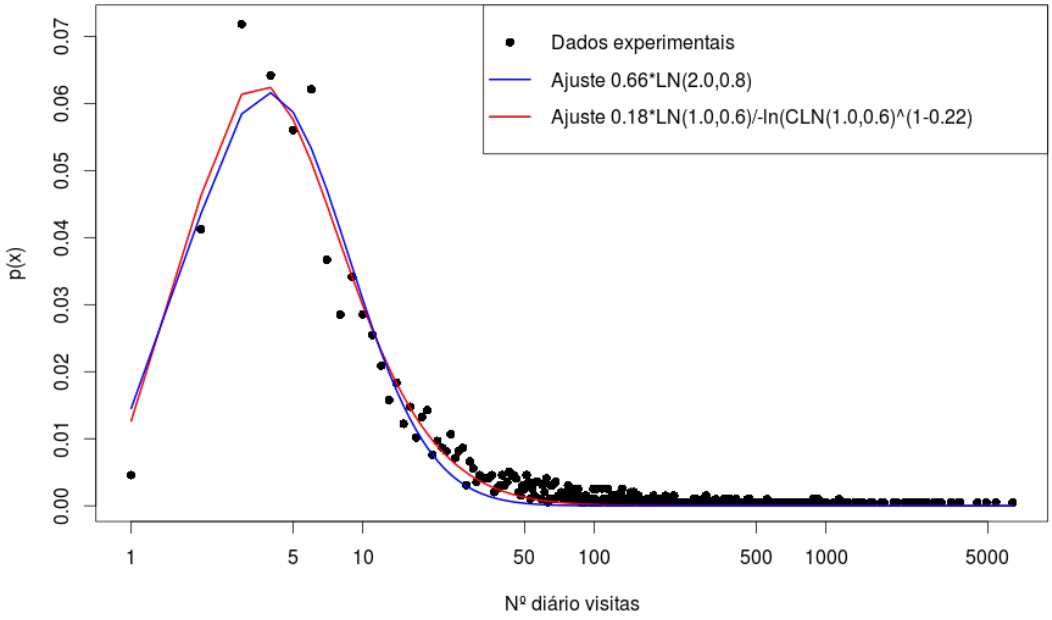


Figura 5.2: Função distribuição da popularidade das visitas, pela média diária, das páginas de um conjunto de 1963 cantores-compositores americanos. Ajuste à equação 4.25 e equação Lognormal com os parâmetros especificados na legenda. Escala linear no eixo das ordenadas.

5.2 Ajustamento do modelo a série de vídeos da rede YouTube

Efectuámos o mesmo teste para um conjunto de dados maior. Desta vez trata-se de um conjunto de 1,687,506 videos da rede YouTube na categoria de entretenimento, em que os dados foram recolhidos através da API da plataforma Cha et al. [2007]. Como podemos observar no gráfico da figura 5.3 há um ajustamento quase perfeito a uma curva lognormal $\ln N(5.516, 2.112)$ a partir do primeiro valor ($Pi_{min} = 1$). Podemos observar no entanto um desvio à curva para valores elevados de popularidade. Fazendo uma ajustamento à equação 4.25 ganhámos uma ordem de grandeza de precisão. Na figura 5.4 podemos observar os ajustes nos quais a equação com o estiramento permite adequar melhor a curva aos valores com uma soma quadrática de resíduos de $9.6x10^{-6}$ por comparação da curva lognormal simples $4.3x10^{-5}$. O erro padrão dos resíduos foi de $3.22x10^{-5}$ em 41244 graus de liberdade. Tendo em atenção a razão $\rho/\lambda = 3.646$ verificamos que se adequa às características da rede Youtube onde de facto todos os dias são adicionados novos videos com uma popularidade muito reduzida.

Outro conjunto de dados é constituido pelo descritivo do número de visitas a um conjunto de 94,282 Videos da rede YouTube na categoria de Ciência e Tecnologia. Na figura 5.5 podemos observar dois ajustamentos efectuados procurando minimizar a estatística de Kolmogorov-Smirnov para diferentes valores do número de pontos ajustados (Pi_{min}). De novo observamos, em comparação com outras funções, que o ajustamento à função lognormal é o mais adequado (KS=0.0069) e que neste caso a função está extendida para valores elevados de Pi, como é previsto no modelo.

Na figura 5.6 está representada a acomodação dos mesmos dados a partir da função densidade de probabilidade. A soma dos residuos quadráticos é de $6.73x10^{-5}$ para o caso do modelo proposto e de $7.63e^{-5}$ para o ajustamento lognormal simples. O erro padrão dos residuos foi respectivamente $8.33e^{-5}$ e $8.86e^{-5}$. Neste caso não se notam diferenças substanciais entre os dois ajustes já que o valor de $\rho/\lambda = 0.963 \approx 1$.

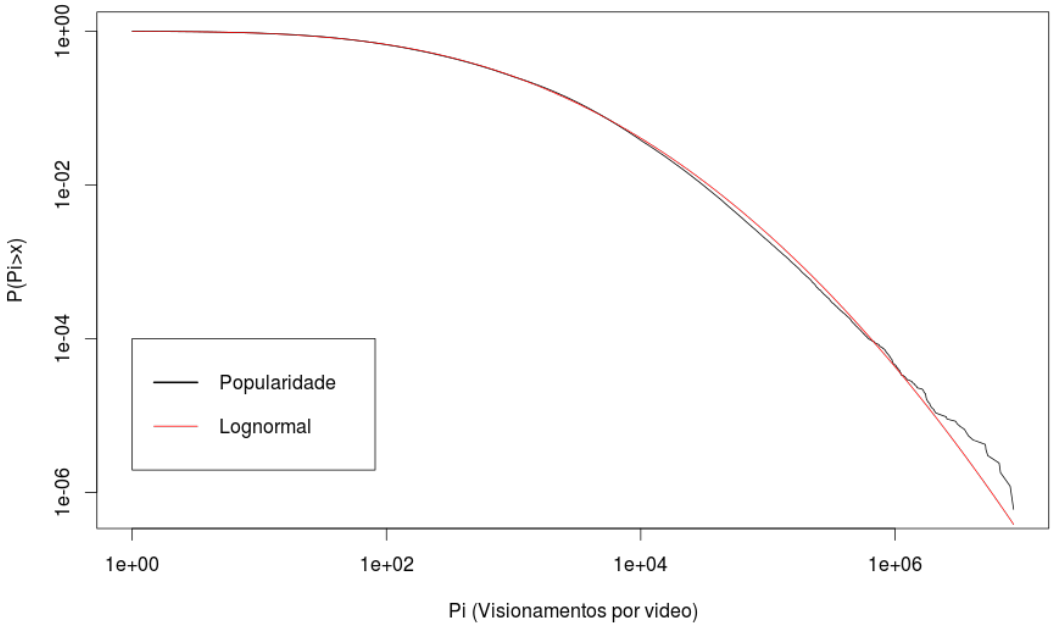


Figura 5.3: Função distribuição acumulada da popularidade do visionamento de videos na categoria de entretenimento na rede YouTube. Valores experimentais e ajustamento lognormal.

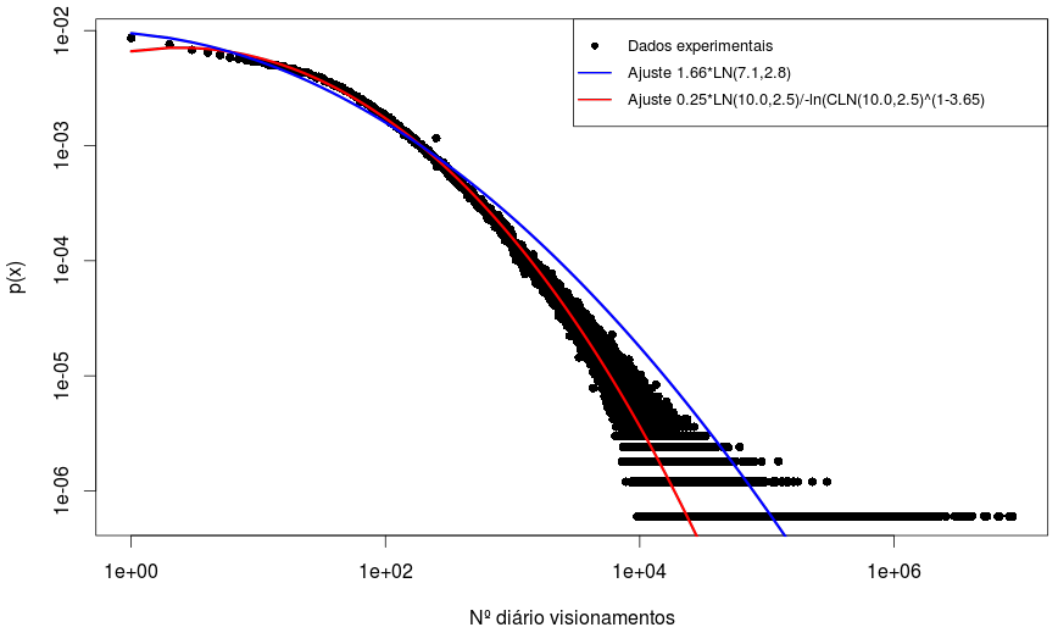


Figura 5.4: Função popularidade do visionamento de videos na categoria de entretenimento na rede YouTube. Valores experimentais e os ajustamentos conforme as formulas da legenda.

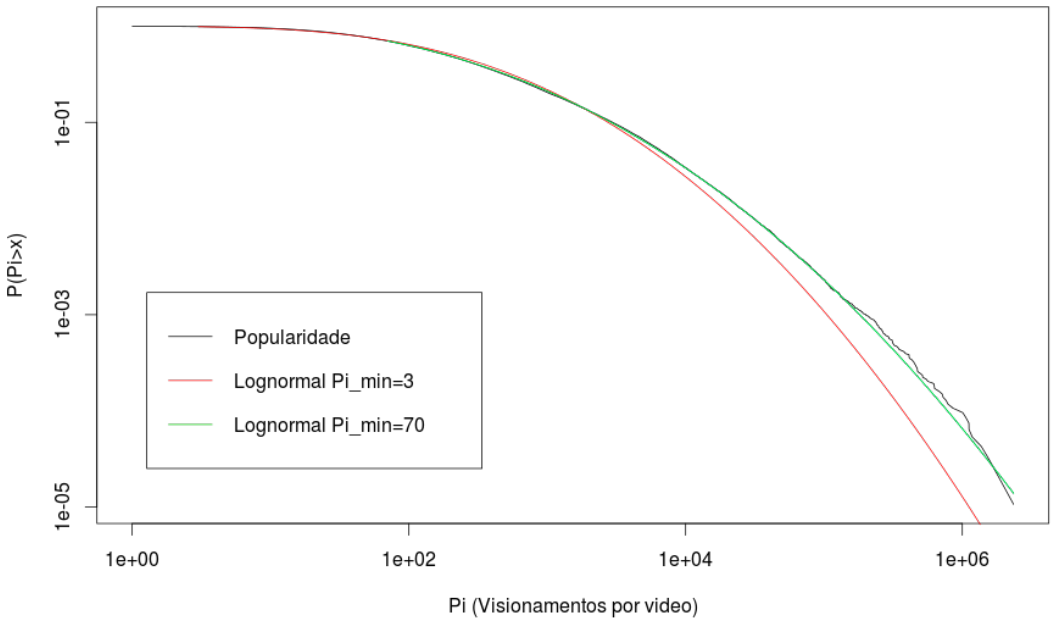


Figura 5.5: Função distribuição acumulada da popularidade do visionamento de videos na categoria de ciência e tecnologia na rede YouTube. Valores experimentais e dois ajustamentos lognormais minimizantes de KS para valores diferentes do conjunto total de pontos determinado por Pi_{min}.

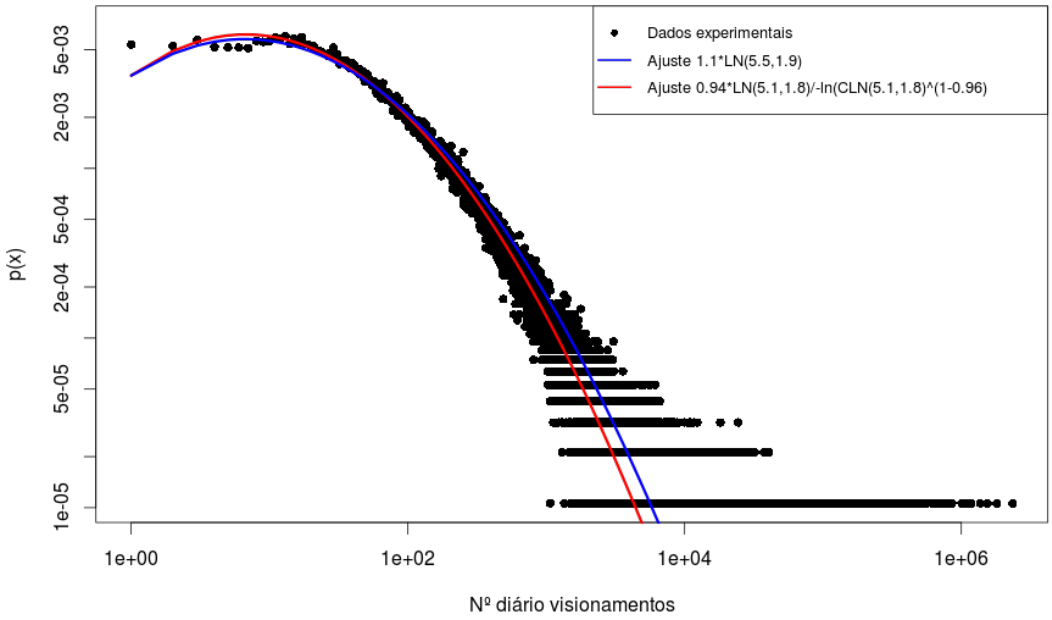


Figura 5.6: Função distribuição da popularidade do visionamento de videos na categoria de ciência e tecnologia na rede YouTube.

5.3 Ajustamento do modelo a série de páginas da Wikipedia

Finalmente analisamos dois conjuntos de dados recolhidos na Wikipedia, respeitantes a visitas a páginas, que apesar de denotarem a popularidade num espaço de tempo relativamente curto dizem respeito a entidades cronológicamente datadas. Trata-se de dois conjuntos, um de albuns de música outro de filmes, lançados no mercado separados por intervalos de décadas.

O primeiro conjunto diz respeito a 728 albuns de música listados na tabela Billboard, desde a década de 1940 até ao ano de 2006, uma vez que as estatísticas das visitas englobam o ano de 2007, com a média de consultas por dia às páginas dos albuns. Encontra-se representado na figura 5.7.

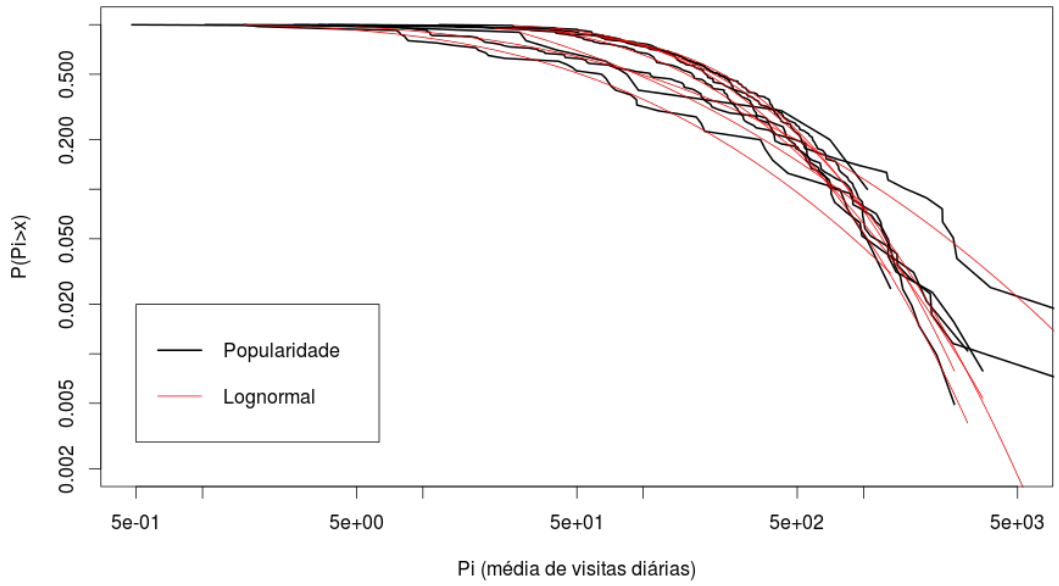


Figura 5.7: Função distribuição acumulada da popularidade das visitas a páginas na Wikipedia respeitantes a albuns da tabelas Billboard em séries temporais correspondentes a diferentes décadas, desde 1946 até 2006. Função complementar cumulativa da distribuição e ajuste a uma função lognormal.

Dado que se trata de um conjunto de dados pequeno, o ajustamento às curvas teóricas não é tão forte como acontece nas séries anteriores. No entanto é significativo que mesmo dentro de parcelas temporais distintas retiradas do conjunto se encontre uma tendência para o ajustamento ao modelo teórico.

Na figura 5.8 vemos representado o ajuste à função densidade de probabilidade para os mesmos dados agora agregados numa série única. Neste caso a soma dos residuos quadráticos foi de $6.42x10^{-4}$ para o modelo e de $6.93x10^{-4}$ havendo apenas uma ligeira melhoria no modelo compensado que propusemos. Este modelo favorece a probabilidade das visitas a páginas menos populares. Tratando-se de dados históricos este facto faz sentido.

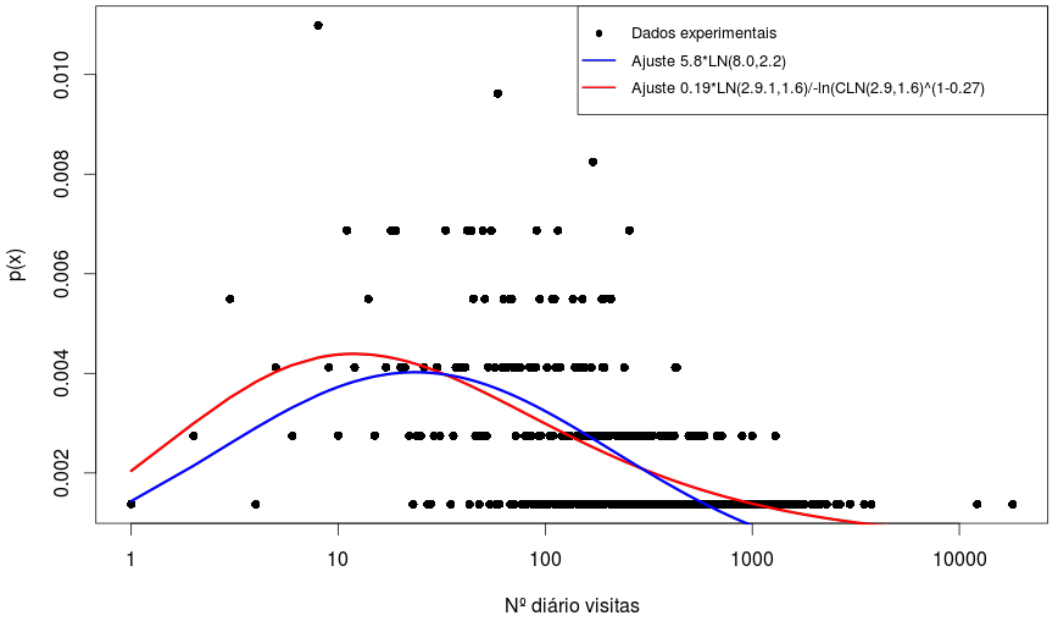


Figura 5.8: Função densidade de probabilidade da popularidade das visitas a páginas na Wikipedia respeitantes a albuns da tabelas Billboard numa série única

O segundo conjunto também se encontra separado em décadas e diz respeito a 1780 filmes lançados nos Estados Unidos nos anos terminados em 6 desde 1936. Também estes dados foram recolhidos da Wikipedia e representam a média de

visitas por dia.

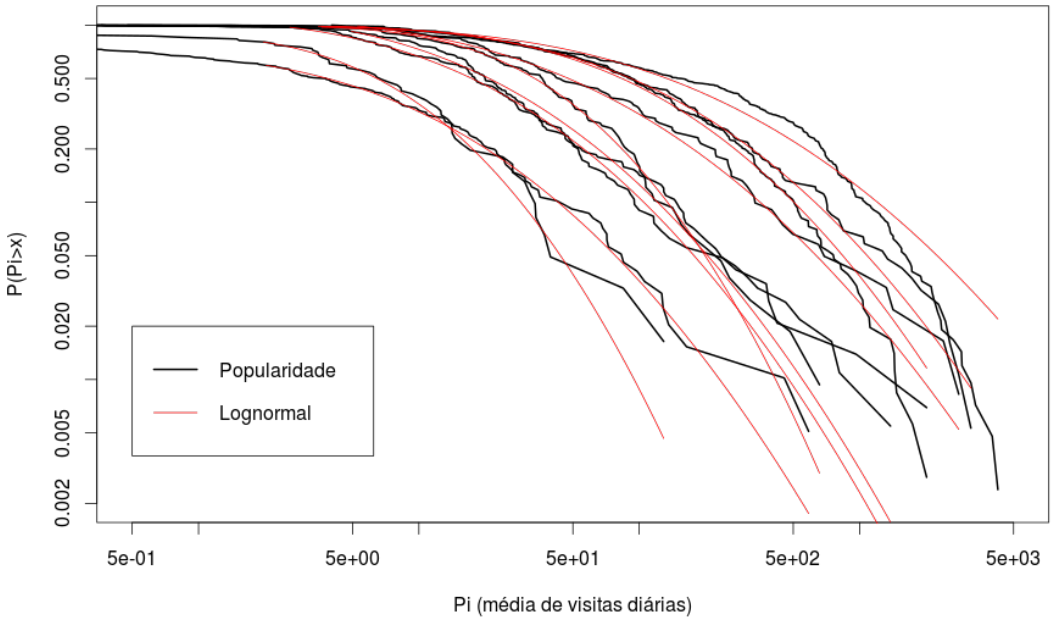


Figura 5.9: Função distribuição acumulada da popularidade das visitas as páginas da Wikipedia respeitantes a filmes lançados nos Estados Unidos em séries temporais correspondentes a anos terminados em 6, desde 1926 até 2006. Função complementar cumulativa da distribuição e ajuste a uma função lognormal.

Nas tabelas 5.2 e 5.3 estão representadas as médias e os desvios padrões de ajustamentos tentativos usando o método adiantado por Clauset et al. Clauset et al. [2009] após 1000 iterações. Da figura 5.9 e das tabelas podemos observar que existe um bom ajustamento aos dados recolhidos mas ressalta que embora o desvio padrão das curvas lognormais se mantém razoavelmente estável, existe um crescimento progressivo das médias de $h'(x)$ com a evolução cronológica da data dos filmes. Curiosamente este facto, não revelado nos albuns da tabela Billboard, revela que os filmes são sujeitos a um esquecimento maior do que as músicas.

Na figura 5.10 vemos representado os mesmos ajustes para o caso da função densidade de probabilidade dos vários anos agregados. Do mesmo modo não se encontram diferenças significativas em torno da média de 3.5 visitas diárias

Ano	KS	Pi_{min}	μ	σ
2006	0.08709073	3.609	4.900122	1.714619
1996	0.05304660	3.495	4.426206	1.402474
1986	0.04881613	4.241	4.523748	1.498674
1976	0.06279041	5.266	3.645491	1.654550
1966	0.06443554	3.621	3.521929	1.079177
1956	0.08008920	4.937	2.744638	1.492465
1946	0.06530308	2.595	2.676155	1.476842
1936	0.05715572	2.027	2.118739	1.512316
1926	0.09072381	1.966	2.002423	1.113024

Tabela 5.2: Parametros do ajustamento das series de filmes. Médias após 1000 iterações de minimização das estatísticas de Kolmogorov-Smirnov.

Ano	KS	Pi_{min}	μ	σ
2006	0.011465237	1.8405180	0.11903784	0.07167119
1996	0.009763638	1.8042257	0.08343909	0.05779264
1986	0.010513925	1.9106464	0.13674311	0.09984284
1976	0.012577459	1.4810258	0.31883068	0.19715906
1966	0.015845422	1.7088825	0.15490175	0.11286822
1956	0.014439695	1.3553110	0.30987914	0.19296964
1946	0.012757817	0.8658665	0.35411985	0.23366274
1936	0.011669109	1.5565005	0.64477716	0.26058973
1926	0.018361456	1.3298920	0.76849105	0.26873439

Tabela 5.3: Parametros do ajustamento das series de filmes. Desvios padrões após 1000 iterações de minimização das estatísticas de Kolmogorov-Smirnov.

(1966). A soma dos residuos foram respectivamente $2.5x10^{-4}$ e $2.9x10^{-4}$ com um erro padrão dos residuos de 0.0022 em 497 graus de liberdade. É notório um desvio de maior probabilidade de visita para os filmes mais populares.

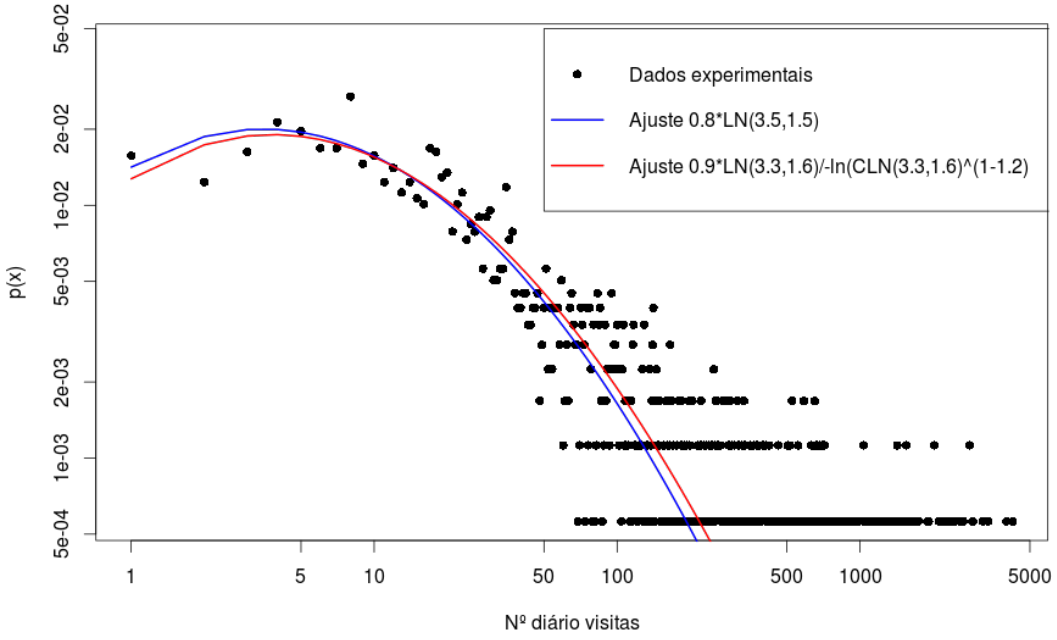


Figura 5.10: Função densidade de popularidade das visitas as páginas da Wikipedia respeitantes a filmes lançados nos Estados Unidos de forma agregada.

5.4 Conclusão

AOs trabalhos de validação que apresentámos neste capítulo mostram que o modelo se adequa bem aos dados experimentais. De facto, nos ajustamentos calculados a equação 4.25 mostrou modelar melhor os dados do que uma equação lognormal simples. Se no caso do ajuste à curva da distribuição completar acumulada a função lognormal representa bem os dados, por comparação com outras funções de distribuição, o ajustamento pela equação lognormal modificada do nosso modelo é melhor em todos os exemplos.

Este modelo simples que analisámos implica um crescimento exponencial da popularidade ao longo do tempo, modulado por uma constante de difusão das mensagens que admitimos ser aleatória segundo o processo de Gibrat. O modelo no entanto não efetua nenhuma previsão para a dinâmica desta constante. A seguir analisaremos um modelo que pretende examinar uma possível formulação para a evolução temporal da γ_i em que analisamos a evolução da popularidade de uma mensagem ao longo do tempo.

Capítulo 6

Modelos Dinâmicos de Popularidade

O modelo proposto nos capítulos anteriores diz respeito à distribuição de popularidade num dado instante temporal por diversas entidades. O modelo foi validado com assuntos consultados nas páginas da Wikipedia e vídeos na plataforma Youtube. O modelo é genérico e pode ser adaptado a outras realidades, no entanto não retrata de uma forma dinâmica como a popularidade pode ser formada. Nos modelos que propomos a seguir a formação da popularidade é explicada pela sua evolução ao longo do tempo. Ambos os modelos são suportados pela suposição de que há um efeito multiplicativo, como foi expresso na equação 4.5 que aqui reproduzimos:

$$\frac{dP_i}{dt} = \gamma_i P_i \tag{6.1}$$

O processo de multiplicação a partir da quantidade de popularidade anterior distingue os dois modelos. No Modelo de Ramificação a multiplicação é efetuada através de ramos ou cascatas de propagação de individuo para individuo. Nesse caso a topologia da rede de contactos entre indivíduos pode ser incorporada, sofisticando o modelo. No Modelo Epidémico a topologia da propagação não é tida em conta, apenas se admite que um determinado número de indivíduos infecta quaisquer outros a uma dada taxa. Não optámos pela incorporação desses parâmetros topológicos no modelo, uma vez que não teríamos forma de validar

esses parâmetros. A secção final deste capítulo analisa o traço comum entre os dois modelos.

Seguidamente detalhamos o Modelo de Ramificação e na secção seguinte o Modelo Epidémico. Posteriormente será discutido o que há de comum entre os dois.

6.1 Modelo de ramificação

Retornando à formulação proposta na equação 4.5 relativa à propagação da mensagem na sociedade, representada na figura 4.1, uma forma de medir a popularidade da mensagem é interpretando o processo como um processo de ramificação. Assim, supomos que existem diversas gerações na propagação de uma mensagem por um conjunto de indivíduos $j \in \{1 \dots N\}$ nas quais a popularidade, ou o número de indivíduos que receberam a mensagem m na geração n é dado por:

$$P(n) = \sum_{j=1}^{P(n-1)} R_j(n-1) \tag{6.2}$$

Em que $R_j(n)$ é o número de indivíduos aos quais é transmitida a mensagem m na iteração n através do individuo j. Conforme atrás afirmámos, este número $R_j(n)$ depende da probabilidade $\beta_j(n)$ da mensagem ser repetida por j, da probabilidade $\alpha_j(n)$ de haver indivíduos disponíveis para a escutar e do seu número Q_j potencial, uma vez que falamos de probabilidades. O parâmetro $\alpha_j(n)$, em conjunção com Q_j está associado ao conjunto de relações pessoais através da qual é propagada a mensagem e $\alpha_j(n)Q_j$ determina o número de indivíduos que potencialmente a podem receber. O fator $\beta_j(n)$, por outros lado, está associado à disposição do individuo j de replicar a mensagem.

Suponhamos, sem perda de generalidade, que existe, para além do processo de ramificação, um processo independente de popularidade da mensagem devida por exemplo à referência na comunicação social, ou apenas por mera menção isolada:

$$P(n) = p(n) + \sum_{j=1}^{P(n-1)} R_j(n-1) \tag{6.3}$$

Afim de tornarmos o problema tratável, deixamos cair o índice j e consideramos idênticos todos os indivíduos.

$$P(n) = p(n) + P(n-1)\langle R(n-1)\rangle \tag{6.4}$$

em que:

$$\langle R(n)\rangle = \langle \alpha_j(n)\beta_j(n)Q_j\rangle_j \tag{6.5}$$

é o número médio de replicações que são escutadas de cada mensagem. Fazendo $\eta = \langle \alpha_j Q_j\rangle$ um parâmetro característico da rede de relações entre os indivíduos, este parâmetro passa a caracterizar o potencial número de indivíduos que escutarão a mensagem.

Entretanto, admitindo que os j indivíduos replicam as mensagens a diferentes instantes t_j, vamos supor também que β não é constante e depende da probabilidade de esquecimento θ, tal que as mensagens aparecerão no fluxo de popularidade a um instante posterior t, denotando memória, replicadas com uma probabilidade $\beta(t-t_j)$. Efetuando a média sobre todos os eventos de replicação função de $\beta(t)$ obtemos uma equação de popularidade da mensagem da forma:

$$P(t) = \int_{-\infty}^{t} P(\tau)\beta(t-\tau)d\tau \tag{6.6}$$

A equação 6.4, transposta para o domínio contínuo, fica então:

$$P(t) = p(t) + \eta \int_{-\infty}^{t} P(\tau)\beta(t-\tau)d\tau \tag{6.7}$$

Onde *ratio de ramificação* η é a média das replicações potenciais de mensagens geradas em qualquer das ramificações, que dependerá das condições de receção α_j e dos recetores potenciais Q_j. Deste modo, η depende da topologia das relações na comunidade, da sua densidade, do comportamento social, da influência entre indivíduos e do suporte da comunicação. Numa rede social esparsa η será baixo; numa rede densa, como num órgão de comunicação social, η será alto. Eventualmente $\eta = 0$ e a mensagem deixa de se propagar. Para assegurarmos estacionaridade restringimos $\eta < 1$ e pela definição de probabilidade $\int_0^\infty \beta(t) = 1$.

Para resolver 6.7 recorremos a uma função de Green $k(t)$, a resposta impulsiva

de $P(t)$, definida como a solução desta equação quando $p(t)$ é um impulso centrado na origem $\delta(t)$:

$$k(t) = \delta(t) + \eta \int_{-\infty}^{t} k(\tau)\beta(t-\tau)d\tau \tag{6.8}$$

e

$$P(t) = \int_{-\infty}^{t} p(\tau)k(t-\tau)d\tau \tag{6.9}$$

Tratando-se de uma convolução, a resolução de 6.8 é imediata pela transformada de Laplace:

$$\mathcal{L}(f * g) = F(s)G(s) \tag{6.10}$$

$$\hat{k}(s) = \frac{1}{1 - \eta\hat{\beta}(s)} \tag{6.11}$$

Pela definição da transformada de Laplace, como tínhamos normalizado $\beta(t)$, temos $\hat{\beta}(0) = 1$ e :

$$\int_{0}^{\infty} k(t) = \frac{1}{1-\eta} \tag{6.12}$$

Ou seja a popularidade média despoletada por uma única menção num instante inicial é igual a $\frac{1}{1-\eta}$. Este resultado pode ser obtido de outra forma, se considerarmos o processo de ramificação no qual cada iteração η mensagens médias são escutadas e replicadas, eventualmente o número de mensagens mencionadas, $t \to \infty$, virá a ser $\sum_{n=0}^{\infty} \eta^n = \frac{1}{1-\eta}$.

A determinação de $P(t)$ vem assim pela função $\beta(t)$. Vamos supor que esta função corresponde a um decaimento exponencial $\beta(t) = e^{-\theta t}$, onde θ corresponde ao factor referido atrás correspondente ao grau de esquecimento da mensagem. De 6.11 a solução para $k(t)$ vem:

$$k(t) = \eta e^{(\eta-\theta)t} \tag{6.13}$$

E a popularidade $P(t)$ vem dada pela integração da equação 6.9.

Consideremos então um caso em que existe o despoletar súbito de popularidade num instante $t = 0$, correspondente por exemplo a um acontecimento acidental ou a qualquer outro fenómeno mediático muito superior ao ruído médio de menções $p(t)$ existente na comunidade. Uma forma de interpretarmos este fenómeno é supormos uma popularidade $P_0\delta(t)$ ocorrida no instante $t = 0$ tal que para $t > 0$:

$$P(t) = P_0 k(t) + \int_{-\infty}^{t} p(\tau)k(t-\tau)d\tau \tag{6.14}$$

Sendo $p(t)$ um processo aleatório, o valor expectável de $P(t)$ vem dado por:

$$E[P(t)] = P_0 k(t) + \frac{\langle p(t)\rangle}{1-\eta} \tag{6.15}$$

$$= P_0 \eta e^{(\eta-\theta)t} + \frac{\langle p(t)\rangle}{1-\eta} \tag{6.16}$$

Onde $\langle p(t)\rangle$ é o ruído médio da mensagem existente na sociedade. Sendo, como atrás vimos, $0 \leq \eta < 1$ para assegurarmos estabilidade de 6.7, é imediato verificarmos que o ruído médio é amplificado pelos factores de densidade das relações de escuta e comunicação na sociedade, sendo eventualmente isolado quando $\eta = 0$.

Por outro lado, a probabilidade das mensagens serem esquecidas, definida no parâmetro θ, funciona em sentido contrário e determina o decaimento da memória da popularidade dos acontecimentos que sobressaem, como em P_0.

Este modelo prevê a resposta a choques situados em instantes $\delta(t - t_0)$ que se podem sobrepor. Não prevê, no entanto, de uma maneira simples, a resposta a uma evolução positiva da probabilidade de replicação. Uma forma de analisarmos este processo é esquecermos o processo de ramificação e admitirmos, como fizemos no modelo de distribuição, uma contaminação generalizada das mensagens como se se tratasse de uma epidemia.

6.2 Modelo Epidémico

Neste modelo mais simples consideramos que existe de igual modo uma função única de replicação $\gamma_i(t)$ para uma mensagem m_i e recuperamos a equação diferencial 4.5 que abordámos no Modelo Estático. Trata-se da equação simples de infetados num modelo epidémico:

$$\frac{dP(t)}{dt} = \gamma(t)P(t) \tag{6.17}$$

No entanto, em vez de escrevermos as equações diferenciais para os diferentes compartimentos da população, vamos concentrar-nos no fator de contaminação $\gamma(t)$ e atribuir-lhe propriedades mínimas:

1. O parâmetro $\gamma(t)$ tem a dimensão t^{-1}, ou seja $\gamma(t) \sim t^{-1}$, uma taxa por unidade de tempo.

2. Numa fase inicial, para que a contaminação da mensagem cresça, já que admitimos não considerar nenhuma popularidade inicial finita, o fator $\gamma(0) \to \infty$.

3. Vamos considerar que γ se mantém finito. Então, decrescendo com o tempo numa fase posterior γ deverá atingir um mínimo em que passará a crescer, ou seja a sua função passa por um ponto de inflexão em que devemos ter $\frac{d^2\gamma(t)}{dt^2} = 0$ num certos instantes $t = L$, ou seja:

$$\frac{d\gamma}{dt} + \gamma^2 = 0 \quad , t = L \tag{6.18}$$

4. Deve haver um instante em que a popularidade atinge um máximo, ou seja, um tempo $t = T_{max}$ no qual $\gamma(t) = 0$ e o factor de crescimento passa a ser negativo.

Os pressupostos 1 a 3 são satisfeitos pela função $\gamma(t) = \frac{c}{t}$. A função analítica mais simples que assegura o quarto pressuposto é obtida tal que:

$$\gamma(t) = -C\frac{\ln(t/T_{max})}{t} \tag{6.19}$$

Aplicando esta função na equação diferencial 6.17 e resolvendo para $P(t)$:

$$P(t) = Ke^{-\frac{C}{2}\ln^2(\frac{t}{T_{max}})} \tag{6.20}$$

Sem perda de generalidade podemos admitir que existem diferentes sementes de epidemia disseminadas pela comunidade, correspondendo a cada uma delas constantes distintas. A popularidade total virá então dada por:

$$P(t) = \sum_j K_j e^{-\frac{C_j}{2}\ln^2(\frac{t}{T^j_{max}})} \tag{6.21}$$

Onde K_j, C_j e T^j_{max} correspondem a constantes distintas de disseminação.

Este modelo ajusta-se na perfeição aos nossos valores experimentais. De facto a equação 6.19, representada na figura 6.1, denota um crescimento do factor multiplicativo da propagação da mensagem afectado inversamente pelo tempo. O fator de crescimento começa por ser muito grande mas positivo, atingindo uma fase em que muda de sinal e provoca um decrescimento da popularidade. As mensagens tendem assim a ter um pico de sucesso replicativo coincidente com o tempo $t = eT_{max}$ que progressivamente se desvanece. Isto independentemente das condições espacias em que se dá a replicação. Neste caso trata-se de replicação generalizada, independente da estrutura da rede de relações entre individuos. As particularidades replicativas individuais de cada mensagem podem ser incluída na constante C_j e as da estrutura de relações que suporta a sua replicação podem ser incluídas na constante K_j que determina em ultima análise a magnitude máxima da popularidade da mensagem. Neste caso, ao contrário do modelo inicial α_j, β_j e θ_j farão parte de C_j.

6.3 Conclusão

Para compararmos os dois modelos, consideremos os parâmetros de replicação determinados no inicio da definição dos modelos: α_i, β_i e θ_i, que denotam as probabilidades da mensagem m_i ser escutada, replicada e esquecida

Se considerarmos a exponenciação da taxa de crescimento instantânea da equação 6.19, que corresponde à solução da equação diferencial linear simples

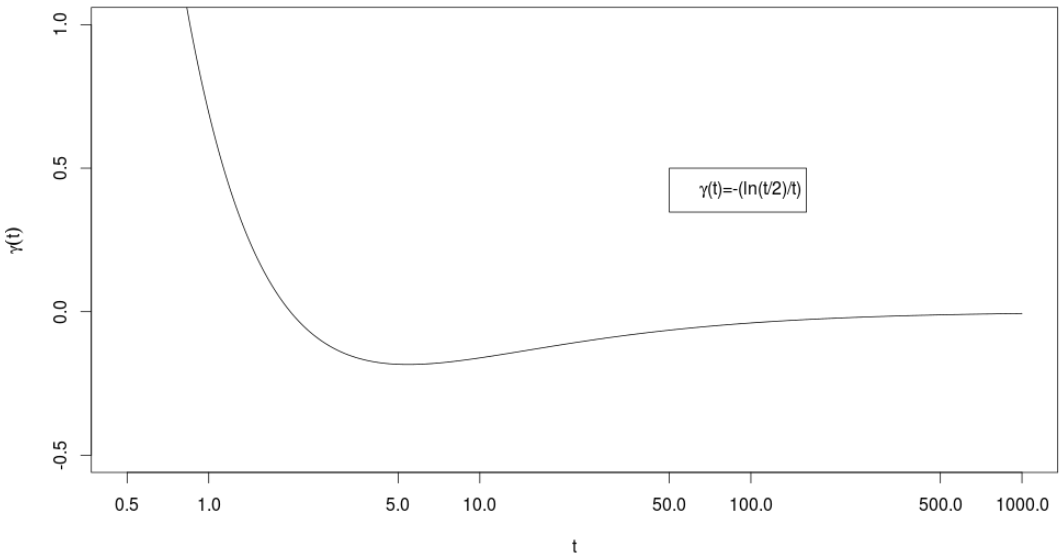


Figura 6.1: A função $\gamma(t)$.

6.17, e suposermos que γ é constante, ou seja que varia a um tempo t' muito mais lento do que t, obtemos a seguinte solução para 6.17:

$$e^{\gamma(t')} = \alpha_i \beta_i t' e^{-\theta_i / t'} \tag{6.22}$$

Onde $\alpha_i \beta_i = 1/T_{max}$ e possui dimensões inversas ao tempo. Este facto faz sentido se pensarmos que o produto dos parâmetros corresponde a uma taxa de replicação por unidade de tempo. Igualmente $\theta_i = C$, com as dimensões do tempo, corresponde a unidades de recordação.

Verificamos em 6.22 que no processo multiplicativo, quer seja o de ramificação quer seja o epidémico, as componentes de taxas replicativas e difusoras e de taxas de esquecimento ou dissipadoras podem ser separadas. A constante $\alpha_i \beta_i$ determina a difusão da mensagem m_i, θ_i determina exponencialmente a magnitude da dissipação.

Finalmente, assinalamos que, fazendo por exemplo $t' = 1$ hora, $\alpha_i \beta_i e^{-} \theta_i = \eta - \theta$ na solução 6.15, menos o termo de ruído comunicacional. Ou seja η representa

bem, como atrás referimos, a parcela da média das replicações até então geradas no intervalo de 1 hora, no processo de ramificação, para um número médio de $\alpha_i\beta_i$ mensagens escutadas e replicadas nessa hora com uma média de θ_i unidades de tempo perdidas nessa mesma hora.

Estes dois modelos agora propostos vão seguidamente ser validados com dados de popularidade evolutivos no tempo. Para o caso de popularidade gerada de forma endógena ao meio vamos tentar ajustar o modelo epidémico às curvas experimentais. Quando o perfil temporal de popularidade assume uma forma abrupta, denotando um impacto externo à comunidade, vamos procurar ajustar o perfil ao modelo de ramificação que contempla impactos exógenos. Ambos os modelos pressupõem um efeito multiplicativo. No entanto, no modelo de ramificação é a função de decaimento $\beta(t) = e^{-\theta t}$, que escolhemos, que melhor se ajusta aos dados experimentais. No caso do modelo epidémico é a função $\gamma(t)$ dada pela equação 6.19 que melhor se ajusta.

Capítulo 7

Validação dos Modelos Dinâmicos de Popularidade

A validação dos modelos vai ser efetuada contra dados experimentais obtidos de trabalhos já publicados anteriormente. Optámos por esta alternativa dada a credibilidade destes dados.

Os dados experimentais que utilizamos são resultado de um processo de tipificação em perfis de evolução temporal obtidos a partir de dados em bruto. O processo está explicado no trabalho a partir do qual são disponibilizados Yang and Leskovec [2011]. Implementámos o algoritmo sobre os dados em bruto e de facto obtivemos os mesmos resultados publicados pelos autores, a partir dos quais fizemos uma pequena alteração, uma vez que encontrámos perfis muito similares no caso dos dados obtidos a partir da rede Twitter. Estes perfis são normativos das evoluções temporais singulares de cada caso e por isso não representam uma cópia fiel de cada um em particular, mas sim a solução "média". No entanto, uma vez que o nosso modelo é também um modelo geral, é razoável ajustar os dois, supondo que neste processo existe uma significância dos resultados. Não são funções ao acaso - os ajustes são efetuados pelo método dos mínimos quadrados e obtêm-se erros nos resíduos muito pequenos, demonstrando a qualidade dos ajustes e a razoabilidade dos modelos.

Na secção seguinte ajustaremos as funções a ambos os conjuntos de perfis e na seção posterior efectuaremos um levantamento das conclusões.

7.1 Teste dos Modelos de Ramificação e de Epidemia com dados de blogues e do Twitter

Em 2009 um grupo de cientistas de Stanford liderado por Jure Leskovec Yang and Leskovec [2011] estudou os padrões de variação temporal das menções de frases-chave e de *hashtags*, em blogues e no Twitter respectivamente, a partir de conjuntos muito grandes (346 milhões de frases e 6 milhões de *hashtags*) recolhidos na Internet. Desses conjuntos separaram um subconjunto dos 1000 mais mencionados e disponibilizaram-nos gratuitamente a título de dados experimentais. Estas duas amostras podem ser agrupadas por padrões típicos de evolução temporal. O trabalho de Yang e Leskovec propõe 6 padrões típicos, tanto para os blogues como para *tweets*. Na nossa investigação encontramos três padrões, no caso do Twitter, cujos parâmetros após o ajustamento são muito próximos. Por esta razão recalculamos os *clusters* segundo o algoritmo dos autores para 4 apenas, e a estes aplicámos o nosso modelo.

Com base na equação 6.21 validámos o modelo epidémico, tentando uma ajustamento de minimização do erro quadrático aos padrões da evolução temporal das menções de *hashtags*. Os ajustamentos aos perfis encontram-se representados nas figuras 7.1 a 7.6.

Um dos padrões típicos acontece quando um determinado *hashtag* é mencionado num espaço curto de tempo e não volta a ser repetido. Tal acontece, por exemplo, quando há um evento mediático como um jogo importante de futebol ou um espetáculo que é tele-visionado. Na figura 7.1 podemos observar o ajuste do nosso modelo a este tipo de evolução da popularidade. A soma do erro quadrático foi de 0.0126 e o erro padrão dos resíduos é baixo: 0.010. O modelo foi ajustado para uma única epidemia, ou seja, pela equação 6.20. $T_{max} = 43$ coincide com a hora de pico, $C = 122.59$ representa a magnitude da taxa de propagação de popularidade e $K = 0.378$ é igual ao valor pico por esta alcançado, normalizado à área do perfil.

Na plataforma Twitter os utilizadores não só colocam conteúdos próprios, como ligam a outros conteúdos através de hiperligações web ou respondem a outros utilizadores, através de menções (*mentions* @), como também replicam conteúdos publicados por outros utilizadores (*retweets* RT). Na tabela 7.1 encontram-

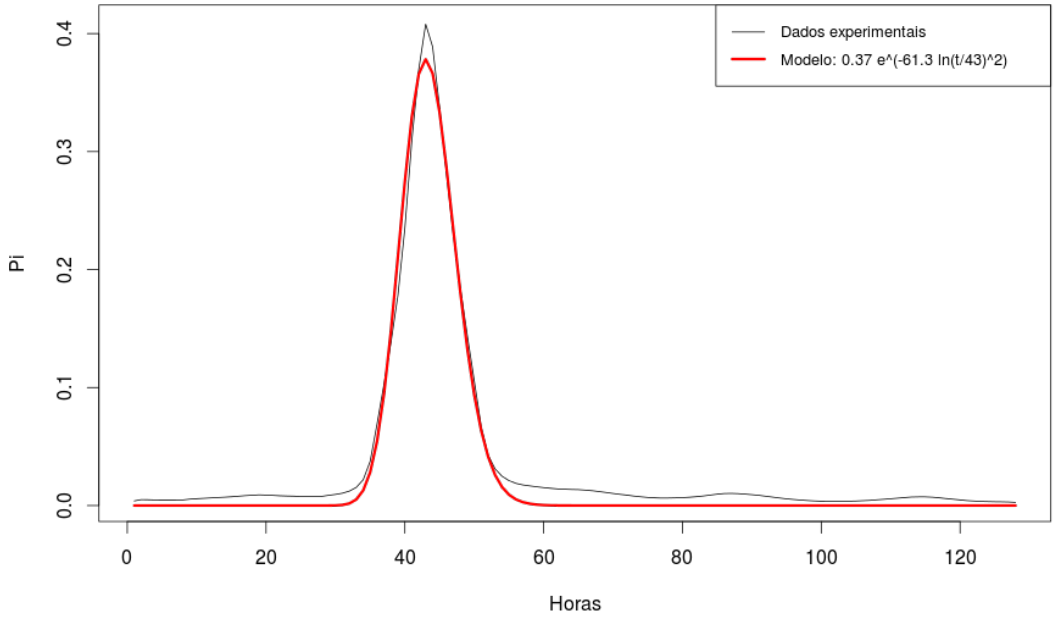


Figura 7.1: Ajustamento (a vermelho) da equação de popularidade epidémica 6.20 a um perfil de popularidade de *hastags* quando a popularidade diz respeito a um único evento.

se listados exemplos típicos para os tipos possíveis de comunicação entre utilizadores no caso exemplificativo de uma conversa à hora de almoço. O símbolo *hiperligação* pode representar uma hiperligação para uma página de um restaurante na Internet e *utilizadorX* o nome de utilizador de uma conta Twitter.

Como podemos verificar o fluxo informativo no seio do grupo de seguidores de um determinado utilizador pode seguir diversos caminhos. Esse fluxo pode ir desde a divulgação de um conteúdo público para todos, até à partilha de um conteúdo privado apenas com outro utilizador.

Tabela 7.1: Modos de Comunicação no Twitter

Formato do tweet	**DE**			**PARA**		
	Mesmo	**Outro**	**Todos**	**Alguns**	**Outro**	**Todos**
hiperligação			X			X
@utilizador2 *hiperligação*			X		X	
@utilizador2 @utilizador3 *hiperligação*			X	X		
RT@utilizador1 Almoço!		X				X
@utilizador2 RT@utilizador1 Almoço!		X			X	
@user2 @user3 RT@user1 Almoço!		X		X		
hiperligação RT@utilizador1 Almoço!		X	X			X
RT@utilizador1 Almoço! *hiperligação*		X	X			X
@utilizador2 *hiperligação* RT@utilizador1 Almoço!		X	X		X	
@utilizador2 RT@utilizador1 Almoço! *hiperligação*		X	X		X	
@utilizador2 @utilizador3 *hiperligação* RT@utilizador1 Almoço!		X	X	X		
@utilizador2 @utilizador3 RT@utilizador1 Almoço! *hiperligação*		X	X	X		
Almoço!	X					X
@utilizador2 Almoço!	X				X	
@utilizador2 @utilizador3 Almoço!	X			X		
Grande restaurante *hiperligação*	X		X			X
@utilizador2 Grande restaurante *hiperligação*	X		X		X	
@utilizador2 @utilizador3 Grande restaurante *hiperligação*	X		X	X		
Eu também RT@utilizador1 Almoço!	X	X				X
@utilizador2 Eu também RT@utilizador1 Almoço!	X	X			X	
@utilizador2 @utilizador3 Eu também RT@utilizador1 Almoço!	X	X		X		
Eu também *hiperligação* RT @utilizador1 Almoço!	X	X	X			X
Eu também RT @utilizador1 Almoço! *hiperligação*	X	X	X			X
@utilizador2 Também vens? *hiperligação* RT @utilizador1 Almoço!	X	X	X		X	
@utilizador2 Também vens? RT @utilizador1 Almoço! *hiperligação*	X	X	X		X	
@utilizador2 @utilizador3 Também vens? *hiperligação* RT @utilizador1 Almoço!	X	X	X	X		
@utilizador2 @utilizador3 Também vens? RT @utilizador1 Almoço! *hiperligação*	X	X	X	X		

Nas figuras 7.2 e 7.3 está representada a magnitude média horária dos fluxos informativos que foram por nós recolhidos na rede social portuguesa que serviu o estudo do Capítulo 5, a partir da API do Twitter, para algumas palavras chave. Apesar das mensagens fluírem a um ritmo acelerado, a maioria do debate dos utilizadores singulares vai crescendo ao longo do dia, concentrando-se no final do dia. Existe um ciclo de utilização da plataforma que a nível horário tem um pico ao final do dia e um crescimento progressivo ao longo deste.

Quando examinados entre fusos horários, os perfis desaparecem. Um outro padrão de popularidade para as menções acontece quando há um único pico, idêntico ao anterior, mas que se prolonga por várias horas num tempo de de-

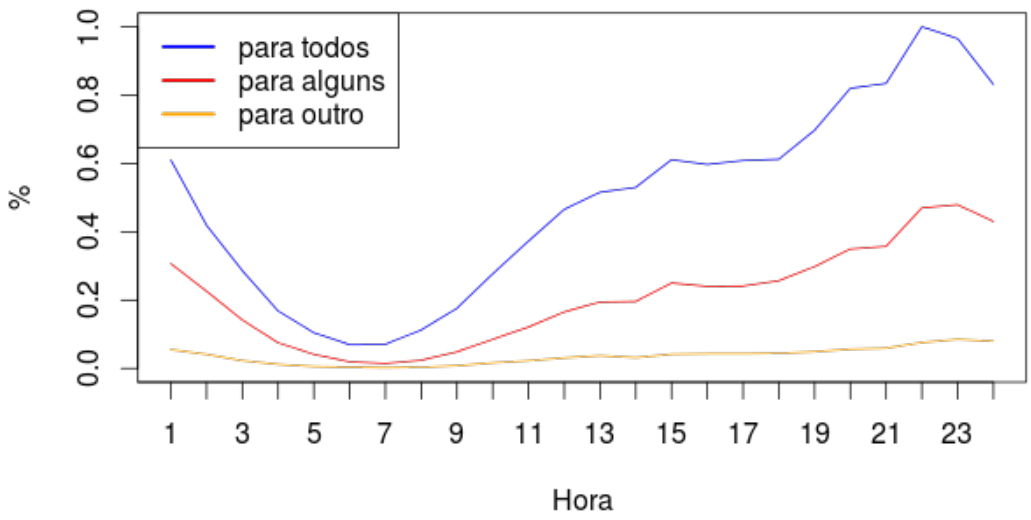


Figura 7.2: Percentagem média das mensagens enviadas diariamente segundo a hora do dia e por modo de envio.

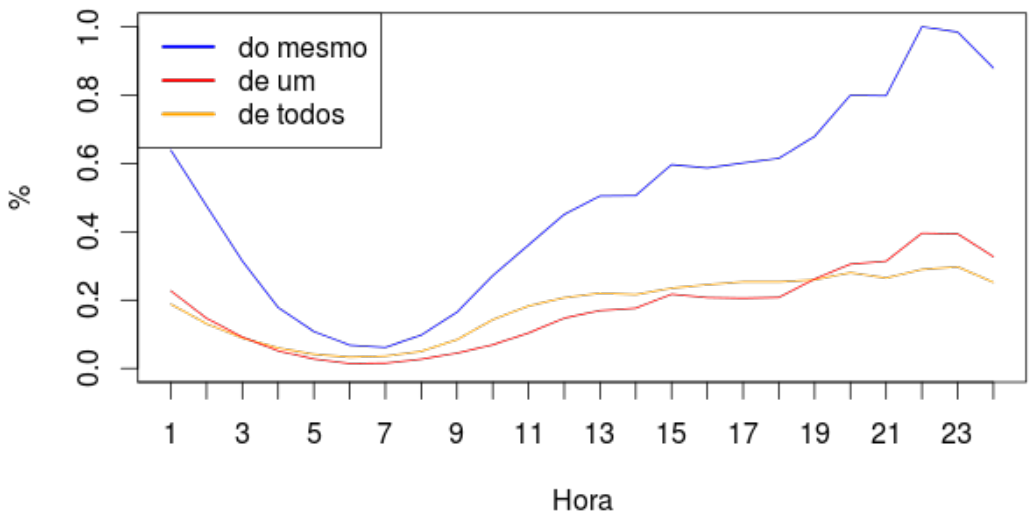


Figura 7.3: Percentagem média das mensagens enviadas diariamente segundo a hora do dia e por origem da informação.

caimento mais lento. Neste caso há uma elevação generalizada da quantidade de menções que se mantém constante ao longo de várias horas. Este caso pode dar-se quando o evento a cuja *hashtag* se refere ultrapassa fusos horários. Na figura 7.4 está representado este padrão obtido através dos dados experimentais. O ajustamento, à semelhança do anterior, é bastante bom, com a soma do erro quadrático igual a 0.00696 e o desvio padrão dos residuos igual a 0.00749. Neste caso o modelo consistiu em duas epidemias com $T^1_{max} = 43$, $K_1 = 0.245$ correspondente à configuração do pico com uma taxa de propagação $C_1 = 79.2$ e uma segunda com um máximo em $T^2_{max} = 49$, 6 horas depois, com magnitude máxima $K_2 = 0.060$ e uma constante de decaimento muito inferior de $C_2 = 5.1$.

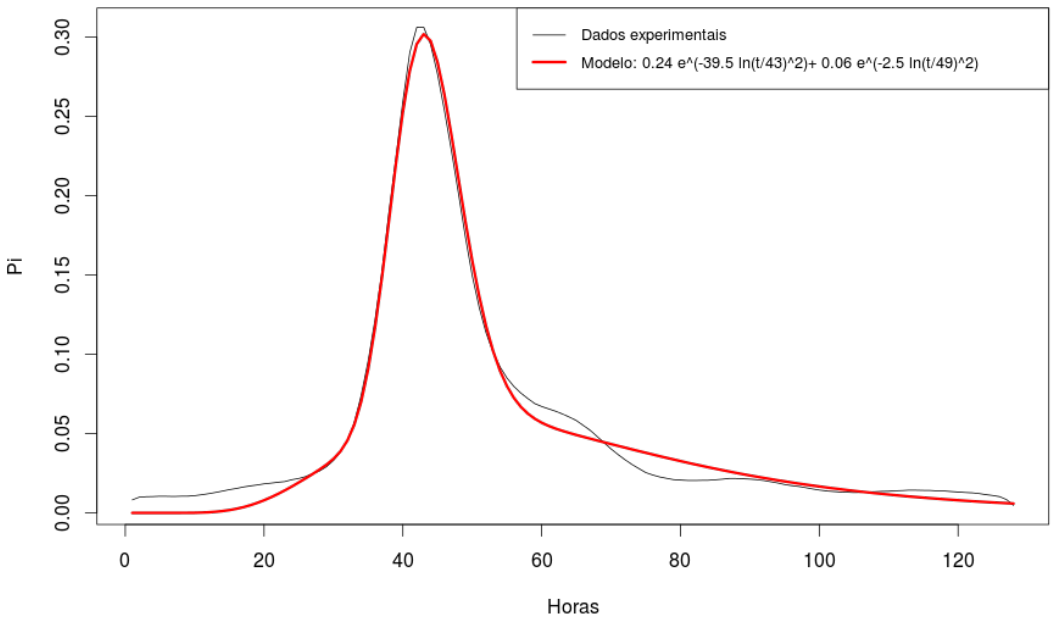


Figura 7.4: Ajustamento (a vermelho) da equação de popularidade epidémica 6.21 ao perfil típico de evolução temporal de menções de *hashtags* quando existe um decaimento lento da popularidade.

Nas figuras 7.5 e 7.6 observamos perfis de popularidade diferentes correspondentes à repetição diária na popularidade do mesmo assunto segundo os ciclos diários referidos na figura 7.2. Na primeira figura 7.5 existe um dia de menção da

mensagem ao qual se segue um pico relevante. Na segunda figura 7.6 as menções após o pico são mais relevantes. Ambos os perfis são ajustados pelo modelo da equação 6.21 com $j \in \{1 \dots 5\}$

Os coeficientes K_j em cada um dos perfis constam na tabela 7.2. Os 4 valores do decaimento,$j \in \{2 \dots 5\}$, ajustam-se muito bem, com uma soma de resíduos inferior a 0.0001, à função:

$$K_j = \frac{(j-1)^{-a}}{4} \tag{7.1}$$

com $a \approx 2$ e $a \approx 1$ respectivamente. Ou seja se encararmos a constante K_j como denotadora da dimensão da rede em que se dá a propagação da mensagem, no primeiro caso ela decresce quadráticamente em cada dia, no segundo, quando o pico despoleta a discusão ela decresce para metade em cada dia que passa.

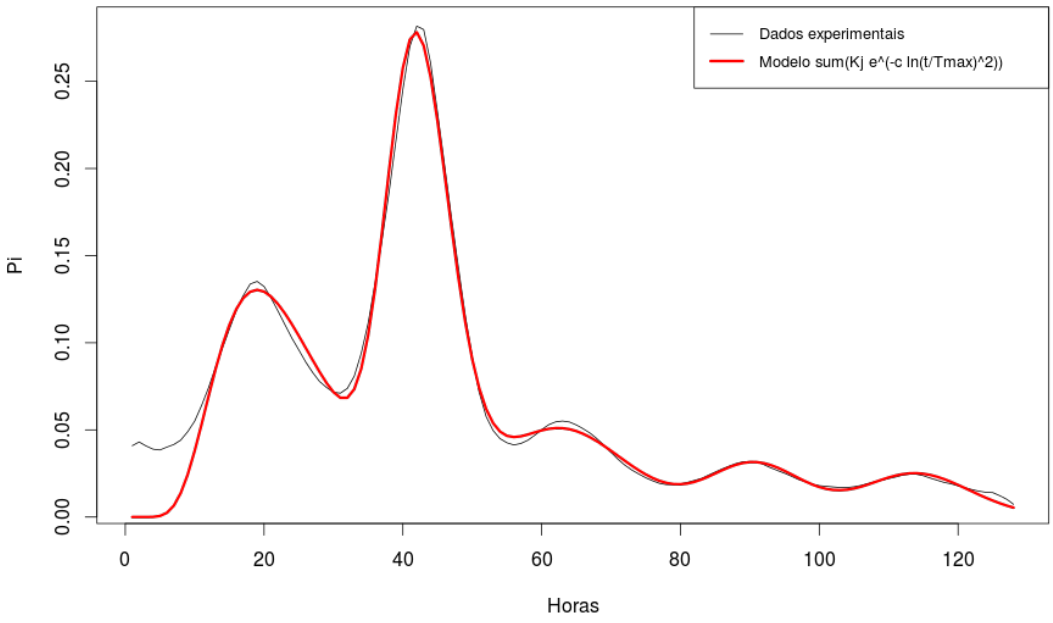


Figura 7.5: Ajustamento da equação de popularidade epidémica 6.21 ao perfil típico de evolução temporal de menções de *hashtags* quando existe repetição.

Os mesmo autores, no âmbito da mesma investigação, disponibilizam também

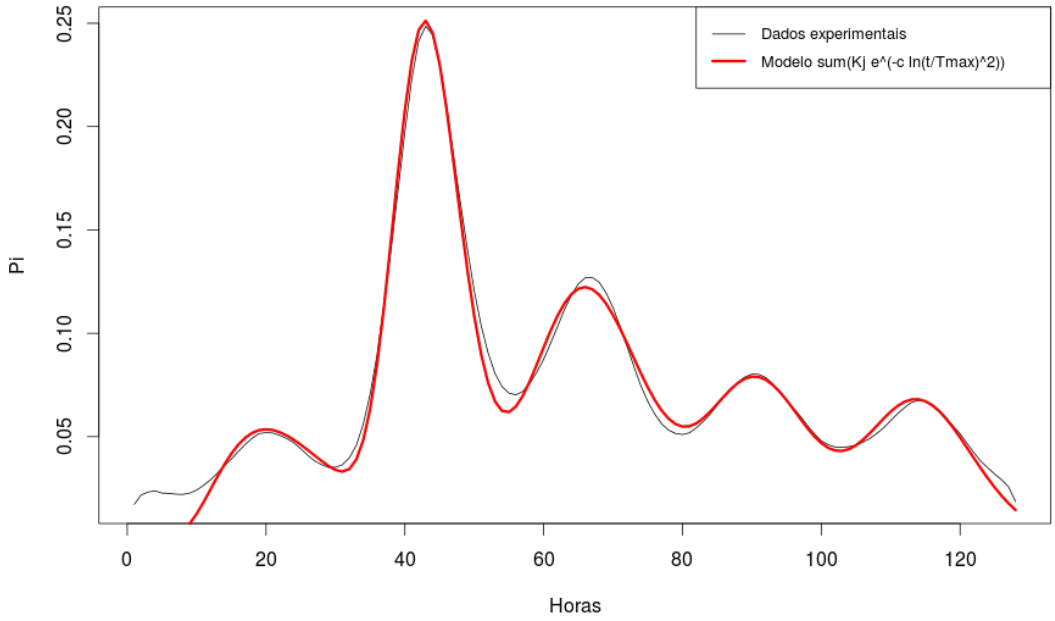


Figura 7.6: Ajustamento da equação de popularidade epidémica 6.21 ao perfil típico de evolução temporal de menções de *hashtags* quando existe repetição.

j	Fig. 7.5	Fig. 7.6
1	0.133644	0.055085
2	0.256918	0.250371
3	0.056868	0.140012
4	0.033230	0.086362
5	0.024687	0.068794

Tabela 7.2: Coeficiente K_j do modelo ajustado ao perfis das figuras 7.5 e 7.6

as series temporais de um conjunto das 1000 maiores propagações de *frase-chave* a que chamam de *memes* num universo de 346 milhões de frases. Estas series temporais foram normalizadas em magnitude e aplicámos o algoritmo proposto no artigo para as agruparmos em 6 perfis típicos de propagação dos *memes* nas horas anteriores e ao longo das horas seguintes ao seu pico de utilização.

Os perfis encontram-se desenhados nas figuras seguintes e ajustados aos modelos que propomos.

No primeiro perfil representado na figura 7.7 podemos observar o ajuste ao modelo de ramificação dado pela equação 6.15 correspondente a um impacto de popularidade em $t = 43$ horas com $P0\eta = 0.5$ e $(\eta - \theta) = 0.112182$, admitimos que o ruído médio das mensagens de fundo foi nulo $\langle p(t) \rangle = 0$. O ajuste foi excelente com uma soma quadrática de residuos de 0.0057468 e um erro padrão de 0.00822 em 82 graus de liberdade.

Segundo este modelo admitimos que a evolução da popularidade não corresponde a uma discussão progressiva e epidémica dentro da comunidade mas que há uma divulgação súbita de uma notícia, ou um qualquer evento, que despoleta uma resposta dos blogues que vai descrescendo em magnitude no tempo.

Efectuámos a mesma acomodação para um perfil ligeiramente diferente em que o pico de popularidade não é iniciado abruptamente e é prolongado algumas horais mais. Conforme examinámos no caso do Twitter (figura 7.4) há um prolongamento da discussão por vários fusos horários devido ao facto do tema extravasar fronteiras. Neste caso o modelo de propagação epidémica ajusta-se muito melhor ao perfil de popularidade. $T^1_{max} = 43$ horas e $T^2_{max} = 53$ horas cerca de dez horas de diferença.

O perfil seguinte diz respeito a um pico único de popularidade cujo ajustamento do nosso modelo é muito bom. A menos uma cauda de resiliência de discussão que o nosso modelo não contempla, os parámetros de $K = 0.333$, $C = 84$ e $T_{max} = 42$ chegaram para ajustar o modelo com uma soma do quadrados dos residuos de 0.0194 e um erro padrão de 0.012 em 126 graus de liberdade.

O perfil seguinte é singular, no entanto foi identificado no artigo original como o perfil mais frequente, correspondente a 28.7% dos perfis analisados. Na figura 7.10 vemos os dois ajustamento efectuados, ambos com um erro nos residuos relativamente baixo de 0.01 em 89 graus de liberdade e um somatório dos quadrados

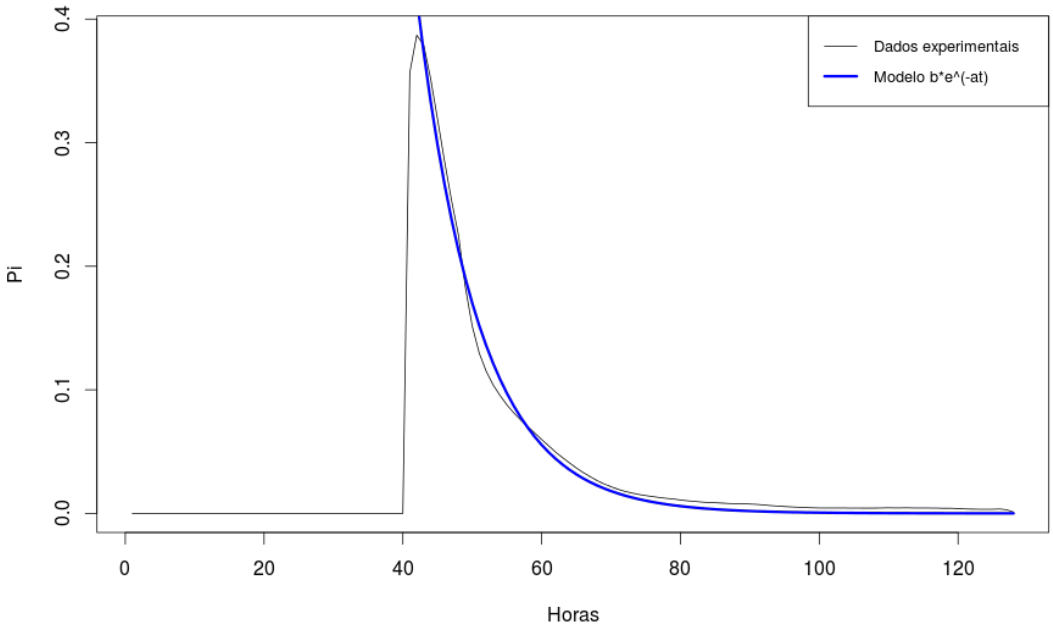


Figura 7.7: Ajustamento da equação de popularidade de memes num dos perfis típicos de propagação em blogues ao modelo dinâmico de ramificação 6.15. O perfil diz respeito à situação em que há um impacto externo à comunidade por uma popularidade abrupta do meme.

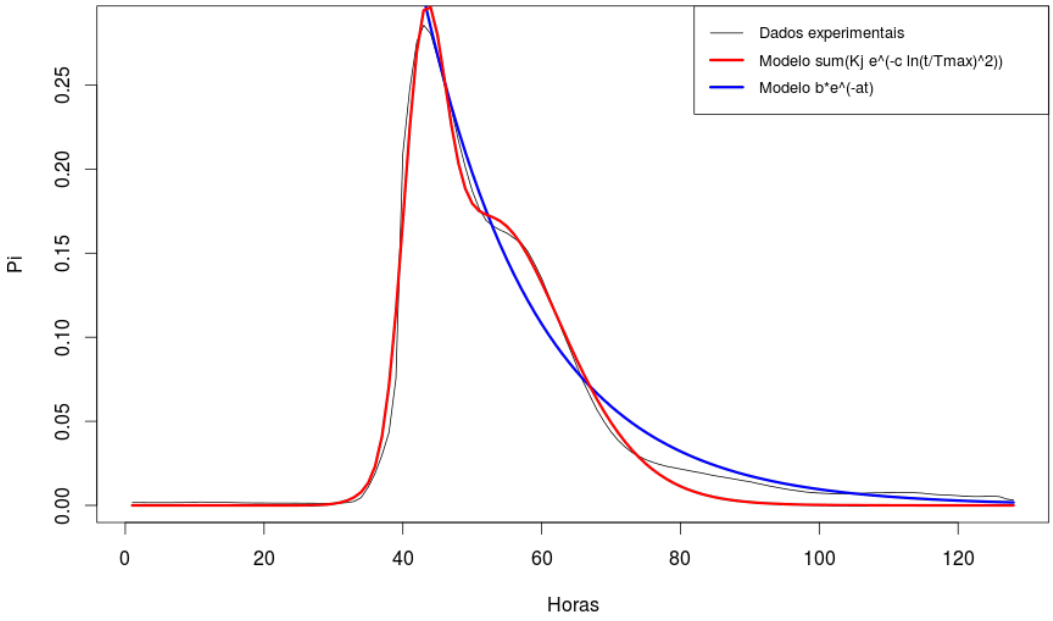


Figura 7.8: Ajustamento da equação de popularidade de memes num dos perfis típicos de propagação em blogues ao modelo dinâmico de ramificação 6.15 e ao modelo epidémico 6.21. O perfil diz respeito à situação em que há um prolongamento da discussão para lá do ciclo diário normal. É equivalente ao perfil da figura 7.4 no caso de propagação no Twitter.

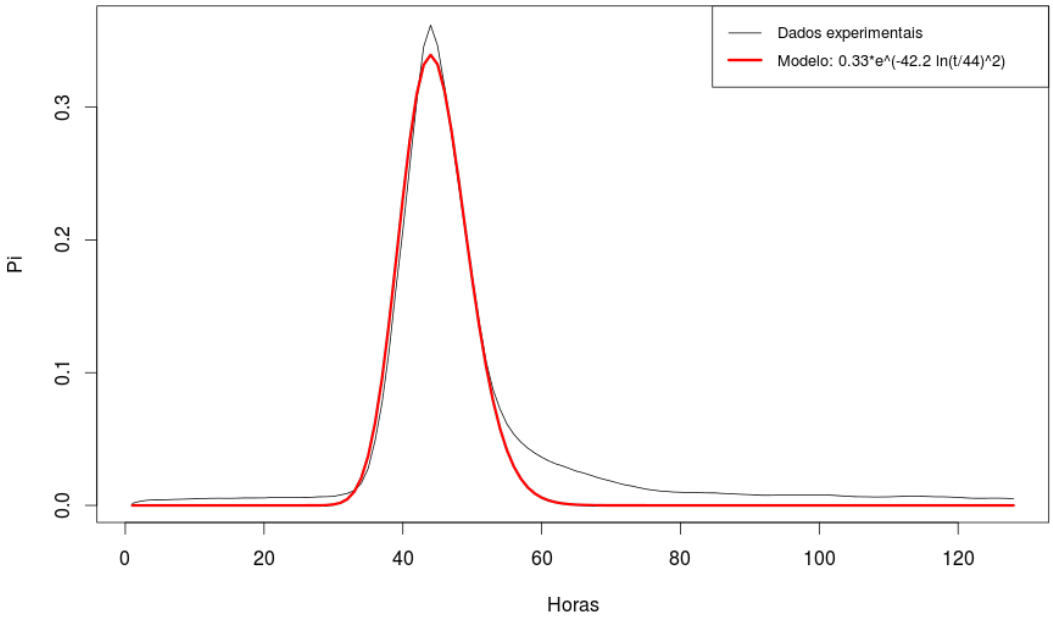


Figura 7.9: Ajustamento da equação de popularidade de memes num dos perfis típicos de propagação em blogues ao modelo epidémico 6.21. O perfil diz respeito à situação em que há um pico acentuado com uma cauda que se desvanece rápidamente. É equivalente ao perfil da figura 7.1 no caso de propagação no Twitter.

de 0.0084 e 0.0076 em 124 graus de liberdade com um somatório de 0.0072. O facto de na cauda descendente do perfil as duas curvas se sobreporem sugere que se trata de processos idênticos. No entanto tal não acontece, pois há a sopreposição de duas epidemias que constroem a face ascendente do pico. De igual modo a aproximação não existe devido ao facto dos processos epidémicos serem similares aos de ramificação. Acontece apenas devido a ter sido especificado um decaimento exponencial em $\beta(t)$ na função 6.7. Um decaimento de outra natureza implicaria mais erros na aproximação da curva, como é patente num decaimento em lei de potências na figura.

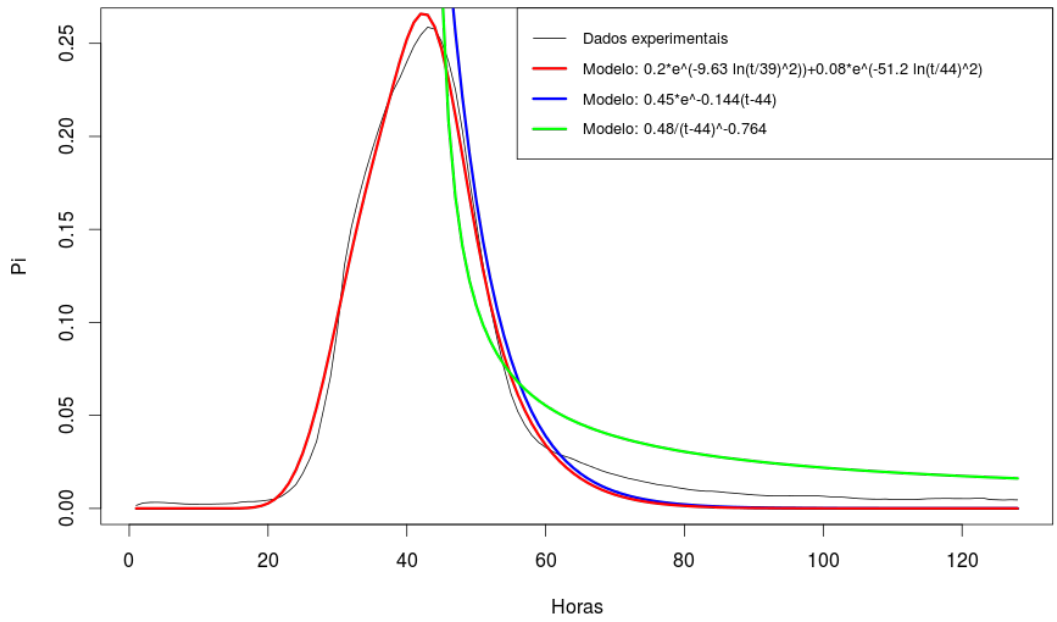


Figura 7.10: Ajustamento da equação de popularidade de memes num dos perfis típicos de propagação em blogues ao modelo epidémico 6.21. O perfil diz respeito à situação em que há um crescimento mais lento da discussão com um desvanecimento rápido. O perfil é ajustado por duas epidemias muito próximas no tempo.

Nas duas figuras seguintes encontram-se os ajustamentos em tudo semelhantes aos das figuras 7.5 e 7.6 no caso dos dados das *hashtags* do Twitter.

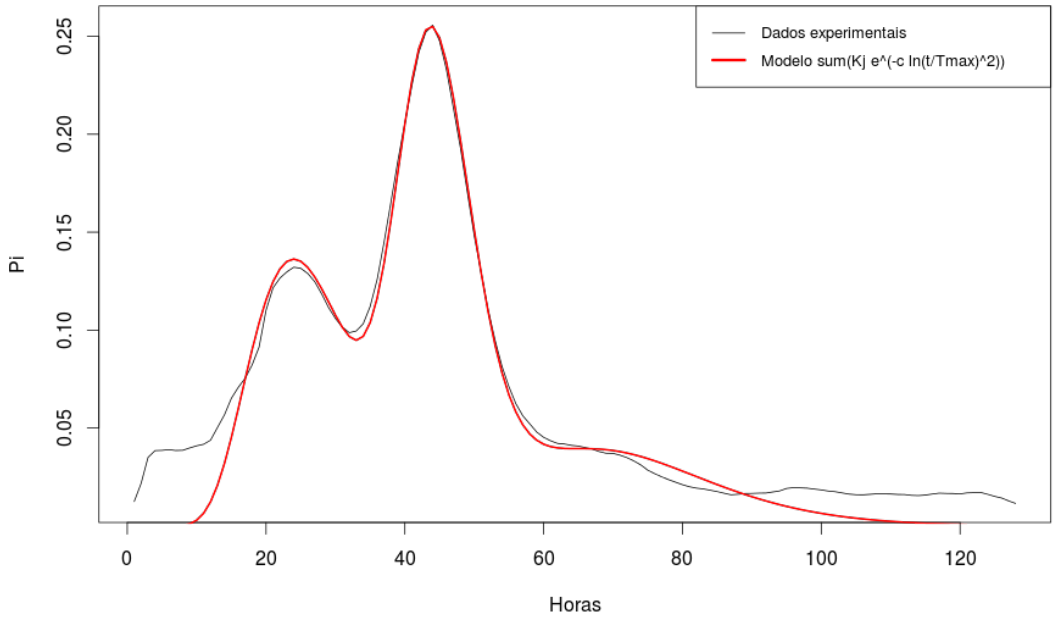


Figura 7.11: Ajustamento da equação de popularidade de memes num dos perfis típicos de propagação em blogues ao modelo ao modelo epidémico 6.21. O perfil diz respeito à situação em que há uma pré discussão que é repetida no dia seguinte, com muito mais polémica, seguindo-se um repetição segundo ciclos diários que decai ao longo do tempo. É equivalente ao perfil da figura 7.5 no caso de propagação no Twitter.

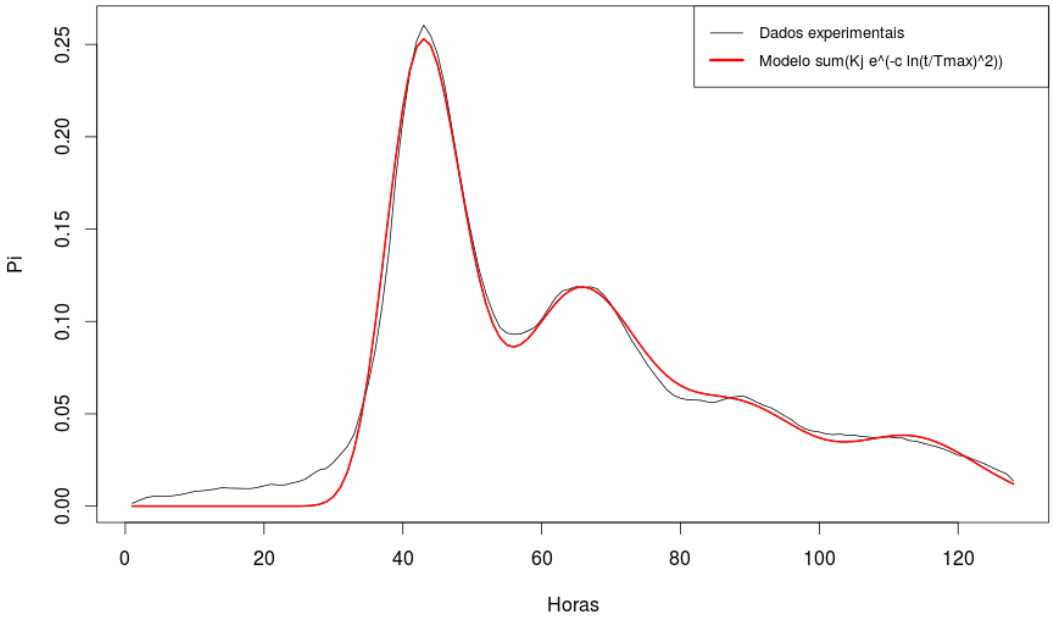


Figura 7.12: Ajustamento da equação de popularidade de memes num dos perfis típicos de propagação em blogues ao modelo epidémico 6.21. O perfil diz respeito à situação em que há uma discussão acesa com um pico acentuado no primeiro dia que é repetida com desvanecimento em dias seguintes. É equivalente ao perfil da figura 7.6 no caso de propagação no Twitter.

7.2 Conclusão

Pudemos verificar na validação dos modelos que, se por um lado os dados experimentais são já em si um modelo dos dados extraídos em bruto das fontes, esse modelo ajusta-se muito bem aos modelo explicativos propostos nesta tese. Se o modelo experimental é um resumo descritivo dos dados experimentais, os modelos que construímos procuram resumir o processo dinâmico de construção da popularidade associada, possuindo um poder explanatório e preditivo. Este tipo de modelos encaixa-se na tipificação exposta por Tatar et al Tatar et al. [2014] no seu resumo de modelos preditivos de conteúdos da Internet, na categoria de análise temporal (*temporal analysis*) semelhante à efetuada por outros autores Pinto et al. [2013]Gursun et al. [2011].

No próximo capítulo abordaremos um caso prático de popularidade, com a análise da influência dos impactos externos dos meios de comunicação social na formação da popularidade, neste caso de candidatos e partidos políticos.

Parte IV

Estudo de Caso

Capítulo 8

Estudo de caso - popularidade e voto

Com exceção do modelo de ramificação, os modelos propostos nesta tese trataram a difusão de informação supondo a inexistência de forças externas a influenciar a propagação da informação no seio dos indivíduos. No modelo com ramificação esse fator é contemplado no termo $\langle p(t) \rangle$ da equação 6.15 que a seguir reproduzimos:

$$E[P(t)] = P_0 \eta e^{(\eta-\theta)t} + \frac{\langle p(t) \rangle}{1-\eta} \tag{8.1}$$

No entanto este termo desempenha no modelo uma influência na resposta a um impacto único, não continuado, na propagação da informação, mas não modela a influência contínua que os vários meios de comunicação efetuam sobre a popularidade de diversas entidades.

Para suprir essa limitação, o estudo de caso que empreendemos e aqui descrevemos procura analisar, partindo de dados experimentais, a influência, não só de uma entidade mas de várias. Foi elaborado um modelo multi-agente que representa a troca de informação no seio da comunidade de indivíduos. Este estudo desenvolve-se sobre os mesmo dados recolhido em duas fases distintas e coletados na rede Twitter durante as eleições presidenciais de Janeiro de 2011 e as eleições legislativas de Junho do mesmo ano, em Portugal.

8.1 Hipótese, objetivo e metodologia do estudo de caso

Em primeiro lugar testámos a hipótese preditiva de resultados eleitorais através da comparação do volume de menções de candidatos e partidos com os valores de sondagens de opinião clássicas e com os resultados finais das eleições. Para este efeito foram recolhidos durante os três meses antecedentes à data das eleições todas as mensagens trocadas entre um grupo de 2000 utilizadores da rede Twitter portuguesa, mais comunicativos em termos de mensagens.

Em segundo lugar, e por duas perspetivas distintas, procurámos compreender os processos de comunicação e de influência que levassem aos resultados obtidos no teste da hipótese preditiva. Para este efeito realizámos simulações multi-agente com dois modelos sociofísicos que nos permitiram analisar esses processos e compreender os fenómenos preditivo e correlativo, quer entre os dados coletados e o resultados final das eleições, quer entre aqueles e os resultados das sondagens clássicas que foram efetuadas na altura.

A conceção da experiência procurou colmatar uma falha há muito apontada na área da sociofísica, que aponta para a falta de validação dos modelos de dinâmica de opiniões [1].

Muitos fenómenos sociais observados à escala macroscópica emergem de comportamentos que obedecem a regras simples a que obedecem um grande número de agentes Schelling [2006]. A descoberta deste fato levou os cientistas sociais à introdução de modelos elementares de comportamento social que o pudesse explicar. Muitos destes modelos assemelham-se a outros introduzidos na moderna física estatística e abrangem áreas tão diversas como a *dinâmica de opiniões*, a *dinâmica cultural*, a *dinâmica da linguagem*, o *comportamento das multidões* ou a *formação de hierarquias* Castellano et al. [2009].

No campo particular da dinâmica de opiniões os modelos mais representativos são os seguintes:

[1] Esta necessidade foi referida por alguns autores, que afirmaram que "Um dos maiores problemas na literatura da 'física social' ou Sociofísica, especialmente na que é dedicada à compreensão dos processos sociais através de simulação computacional, é a da falta de conexão com exemplos reais"Sobkowicz [2009]Moss and Edmonds [2005]

- O modelo Voter Clifford and Sudbury [1973] - cada agente fica com a opinião da maioria dos seus vizinhos.
- O modelo da regra maioritária Galam [2002] - um grupo aleatório de r agentes da comunidade, por turnos, toma a opinião da sua maioria.
- O modelo Sznajd Sznajd-Weron and Sznajd [2000] - um par de agentes vizinhos com valores binários de opinião determina a opinião da sua vizinhança.
- O modelo Deffuant Deffuant et al. [2000] - os agentes, com níveis de opinião definidos por valores contínuos, discutem aos pares e tendem a convergir para um valor de compromisso.
- O modelo Hegselmann-Krause Hegselmann and Krause [2002] - um agente discute com todos os seus vizinhos a sua opinião definida por um valor contínuo e tende a convergir para um valor de compromisso.
- Os modelos Brownianos Schweitzer and Garcia [2010] - Os agentes interagem como partículas brownianas de acordo com uma equação da física estatística.
- A Teoria do Impacto Social Nowak et al. [1990] - cada agente é influenciado pelos seus vizinhos tendo cada um valores próprios de suporte e de influência sobre cada vizinho.

Neste estudo de caso escolhemos os dois últimos modelos listados acima para analisar o processo de formação de opinião na rede e para validar esse processo com os dados de opinião recolhidos na rede Twitter. Neste percurso partimos de dois pontos de vista distintos, correspondentes aos dois níveis de análise ilustrados na figura 8.1 :

- Nível (a) - Por um lado admitimos que a formação de opinião depende da influência dos media e das notícias, mas também da informação que tem a comunidade no seu todo como alvo, como é o caso da informação publicitária. Para este efeito utilizámos o modelo Browniano de Schweitzer

Schweitzer [2007], que é sobretudo centrado na dinâmica temporal de contaminação entre o domínio social e o domínio individual de cada agente. É um modelo baseado no modelo de movimento browniano das partículas dos gases formalizada pela equação de Langevin Langevin [1908] e que recorre à metáfora da partícula, condicionada não só pelas suas características próprias mas por um efeito estocástico a uma escala temporal mais curta, a da temperatura, para a compreensão da formação e da polarização de opiniões numa comunidade de agentes.

- Nível (b) - Por outro lado procurámos analisar o efeito da interação das pessoas para a formação da opinião. Utilizámos um modelo baseado na Teoria do Impacto Social de Latané Latane [1981]. Este modelo centra a sua análise na influência entre indivíduos para a formação da opinião. É portanto mais adequado quando pretendemos analisar os efeitos estruturais das redes de relações na formação coletiva de opinião. No modelo cada individuo tem um certo grau de influência e de suporte sobre a opinião de outro e a formação desta é baseada no efeito agregado da vizinhança entre agentes.

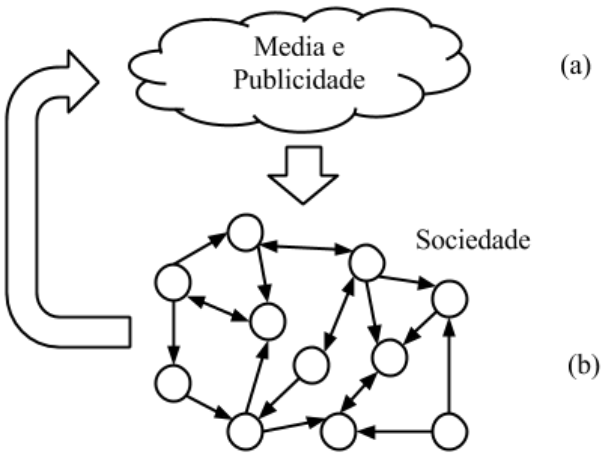


Figura 8.1: Níveis de influência na formação de opinião: (a) influência a partir dos media e publicidade (b) influência a partir da comunicação entre indivíduos

8.2 Dados coletados relativos ao debate eleitoral numa rede social

Os dados que coletamos da rede Twitter foram recolhidos através de dois *scripts* em Python [1] correndo sobre a interface de programação da rede Twitter [2]. Coletamos dados de 1903 utilizadores singulares, dentro dos mais assíduos na rede portuguesa, depois de termos feito uma amostragem do universo de utilizadores, uma vez que a recolha de dados não comercial é quantitativamente limitada pela empresa. Para além dos utilizadores normais coletamos os *tweets* das principais contas de meios de comunicação social como jornais, televisões e rádios. Deste modo separámos os conteúdos entre dois grupos com comportamentos bastante distintos. As contas de comunicação social colocam exclusivamente os títulos das notícias em simultâneo com a sua publicação através dos seus canais de distribuição.

No que respeita aos conteúdos publicados, os utilizadores tendem a expressar opiniões com veemência e quando falam de política tendem a ser frequentemente agressivos e polarizados, confirmando os estudos já efetuados Kirzinger [2011]. Em vez de orientarmos o nosso estudo para uma classificação do conteúdo, conforme foi anteriormente feito por outros autores Tumasjan et al. [2010] O'Connor et al. [2010] Pak and Paroubek [2010] Amigó et al. [2013], optámos por centrar o trabalho na contabilização das menções aos candidatos ou partidos. Em termos de relevância dos resultados esta opção não se mostrou significativamente inferior a outras experiências em que foram implementados algoritmos de desambiguação de entidades e análise de sentimentos Amigó et al. [2013]. Tal deveu-se, do nosso ponto de vista, ao facto do simples nome dos candidatos ou dos partidos ser por si pouco ambíguo e ao volume de sentimento neutro ou mal classificado ser ainda bastante relevante no atual estado da tecnologia. De igual modo, para o propósito desta tese, é mais relevante a mera comunicação de informação associada a uma entidade do que as propriedades semânticas dessa informação.

[1] Python (http://www.python.org/ é uma linguagem de programação muito versátil e facilmente programada para prototipagem de aplicações.

[2] Trata-se da API REST 1.1 da rede Twitter https://api.twitter.com/ que permite recolher não só dados sobre os utilizadores como também os conteúdos que trocam na rede.

Inicialmente mostramos que a quantidade de notícias produzida pelos media, acerca de cada candidato ou partido, acompanhou de perto os resultados finais das eleições e os resultados das sondagens efetuadas durante as pré-campanhas e campanhas eleitorais. Este resultado já tinha sido reportado em 2007 num levantamento feito em jornais e acompanhado noutros trabalhos Véronis [2007] Tumasjan et al. [2010]O'Connor et al. [2010]. Apesar deste resultado, aparentemente paradoxal, não ter uma explicação imediata, acontece adicionalmente ao número de citações pela população em geral - a proporção de referências a cada entidade em concurso numa eleição acompanha em termos relativos, e com uma elevada aproximação em termos absolutos, o resultados final das eleições. A explicação para este resultado prévio orientará o estudo de caso.

8.2.1 Notícias, sondagens e tweets

Na Figura 8.2 está representada a magnitude do número de *tweets* difundidos pela comunicação social respeitante a cada candidato das eleições presidenciais comparativamente às sondagens clássicas que foram sendo divulgadas pelas diversas agências de sondagens e ao resultado eleitoral final que se encontra reportado na Tabela 8.1. Afim de não prejudicar a clareza da leitura a magnitude horária dos *tweets* foi interpolada por uma linha de tendência que aponta a projeção final.

Candidato	Resultado Final
Cavaco Silva	53,14%
Manuel **Alegre**	19,67%
Fernando **Nobre**	14,04%
Francisco **Lopes**	7,05%
Defensor **Moura**	4,52%
Manuel **Coelho**	1,58%

Partido	Resultado Final
PSD	41,19%
PS	30,42%
CDS	12,72%
PCP	8,61%
BE	5,69%

Tabela 8.1: Resultados finais das eleições.

Na Figura 8.3 estão representados os mesmos dados de sondagens mas neste caso a magnitude de tweets representa a magnitude do conteúdos gerados pelo grupo de utilizadores coletados. Na Figura 8.4 e na Figura 8.5 encontra-se a mesma análise desta vez para as eleições legislativas.

Afim de se avaliarmos o grau de similaridade entre as series temporais efetuámos o calculo de uma correlação de Pearson ponderada obedecendo às seguintes fórmula:

$$\langle \rho \rangle = \frac{1}{N} \sum_{i} k_s \frac{E[(T_i - \mu_{T_i})(S_i - \mu_{S_i})]}{\sigma_{T_i} \sigma_{S_i}} \tag{8.2}$$

A fórmula parte do suposto que a intenção de voto determina a influência das referências aos candidatos ou partidos. Nesta fórmula k_s é uma serie de constantes de ponderação arbitrária, cujo propósito é o de penalizar os valores mais antigos antes das eleições. Partimos do pressuposto que a referência às entidades, a partidos ou candidatos, decresce em importância no passado. Afim de comprovarmos esta hipótese efetuamos o cálculo para uma progressão linear, uma progressão harmónica e sem ponderação. A variável T_i representa o vetor com a percentagem de magnitude de *tweets* para o dia i com a dimensão do número de candidatos/partidos e S_i é o vetor equivalente representando as ultimas sondagens antes do dia i e N é o número total de dias recolhidos. Os resultados constam na Tabela 8.2.

Serie	$\langle\rho\rangle$ $k_s = 1.0$	$\langle\rho\rangle$ $k_s = k_{s-1} - s/n$	$\langle\rho\rangle$ $k_s = k_{s-1}/s$
eleições presidenciais media	0.737	0.824	0.866
eleições presidenciais utilizadores	0.750	0.837	0.877
eleições legislativas media	0.857	0.866	0.846
eleições legislativas utilizadores	0.912	0.917	0.925

Tabela 8.2: Coeficiente de correlação de Pearson para as séries temporais sem ponderação ($k_s = 1.0$); com uma ponderação linear $k_s = k_{s-1} - s/n$ onde $s \in \{1 \dots n\}$ representa um período de 7 dias - uma semana entre n semanas; e uma ponderação harmónica ($k_s = k_{s-1}/s$).

Relativamente à correlação de Pearson entre os vetores de *tweets* de utilizadores e media noticiosos, verificam-se os seguintes valores:

- Presidenciais: 0.885
- Legislativas: 0.909

Destes dados examinados podemos observar o seguinte:

- O fluxo de *tweets* dos utilizadores ao longo do tempo está bastante correlacionado com o fluxo noticioso sobre as mesma entidades.
- Quer o fluxo de *tweets* dos utilizadores quer o fluxo de *tweets* dos media, em percentagem relativa, aproximam-se bem da intenção de voto dos eleitores medida nas sondagens e nas urnas.
- Estes resultados são mais verdadeiros nas eleições legislativas. Talvez pelo carácter pessoal das eleições presidenciais pontuado por acontecimentos episódicos o debate não aproxima tão bem a intenção de voto expressa em sondagens e urna.

Afim de avaliarmos a relação entre os fluxos noticiosos e o debate na rede, calculamos a covariância entre estes vetores em função de um deslocamento temporal. Nas figuras 8.6 e 8.7 estão representadas estas medidas. Como podemos verificar, existe um pico de correlação para o próprio dia. Curiosamente podemos observar, no caso dos candidatos presidenciais menos debatidos (Coelho e Moura),

que esse pico é mais pronunciado para o próprio dia dada a menor resiliência do seu debate ao longo do tempo. Os episódios são debatidos no próprio dia e depressa desaparecem da discussão popular. Outra explicação para este fato é que o debate sobre estes candidatos foi quase exclusivamente motivado pelos media. De uma análise transversal sobre os conteúdos podemos verificar que raramente são mencionados fora do âmbito de uma notícia.

Perante este dados e o esquema circular de influência atrás referido na Figura 8.1, é natural a colocação da hipótese da existência de um fenómeno de influência mútua, e nos dois sentidos, entre os conteúdos mediáticos e o debate na população, que resulta num condicionamento efetivo da intenção e na prática do voto.

Poder-se-à supor também que este tipo de influência muito direta coincide com o clássico modelo *hipodérmico* de comunicação de massas Wolf and de Figueiredo [1987], hoje em dia ultrapassado, no qual se admite uma determinação forte, a partir da mensagem dos media e a ação do individuo. Apesar de se tratar de um modelo excessivamente determinista esta influência não deixa de fazer sentido. Mesmo nos modelos de comunicação mais atuais que pressupõem uma reflexão do individuo sobre os conteúdos, que pressupões a influência dessa reflexão sobre os conteúdos ou a hipótese de um *Agenda Setting* a determinar a evolução dos temas, não deixa de fazer sentido que a expressão da sociedade acompanhe o que é expresso nos media. Não se trata de ação mas antes dos conteúdos mentais que são expressos em público. O fenómeno mais interessante, que de fato queremos analisar nesta tese, é que de fato alguma ação dos indivíduos - o seu voto expresso nas eleições - acompanha a magnitude dessa expressão debatida e motivada pelos media. Esta hipótese é confirmada na teoria do *Agenda Setting* McCombs and Shaw [1972] que prevalece como das explicação mais relevantes para os fenómenos de formação de opinião pública. De facto, tal como McCombs e Shaw também este estudo confirma a forte correlação entre os temas debatidos pelo público e os temas lançados pelos media.

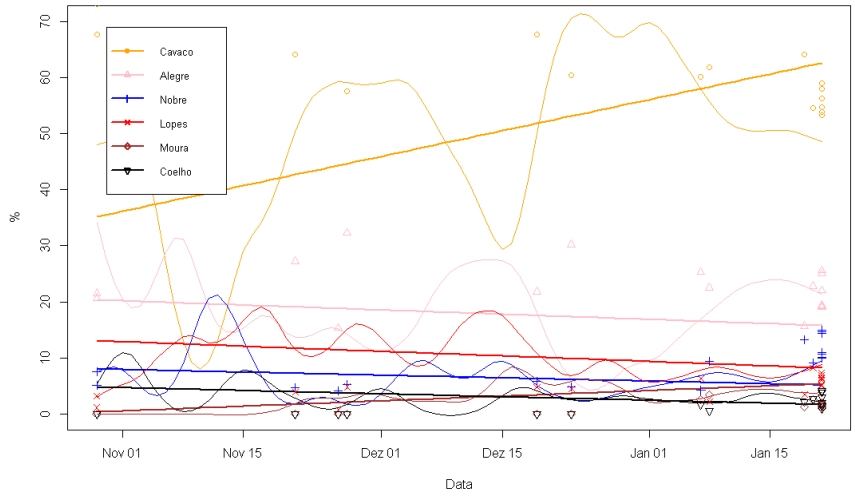


Figura 8.2: Interpolação da percentagem de *tweets* emitidos pelos media referenciando cada candidato presidencial durante a campanha. Linha de tendência de 1ª ordem minimizando o erro quadrático e valor das sondagem representados por símbolos pontuais. Total de 44 contas de media

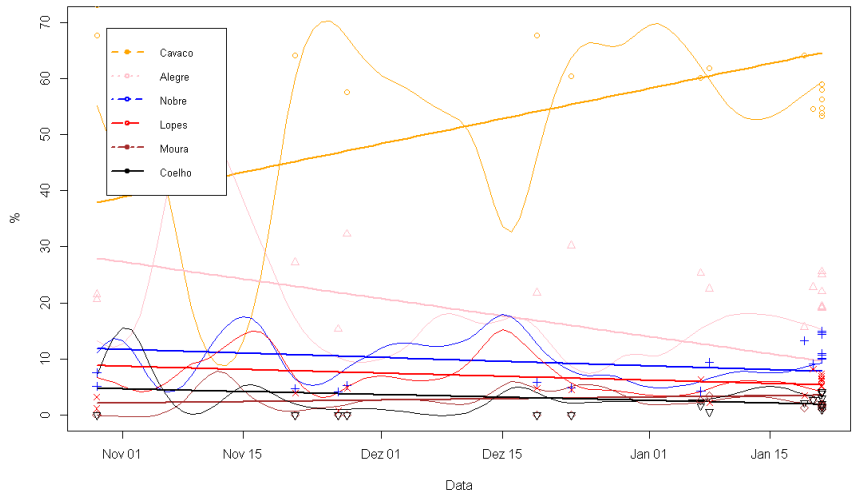


Figura 8.3: Interpolação da percentagem de *tweets* emitidos pelos utilizadores referenciando cada candidato presidencial durante a campanha. Linha de tendência de 1ª ordem minimizando o erro quadrático e valor das sondagem representados por símbolos pontuais. Total de 1903 utilizadores.

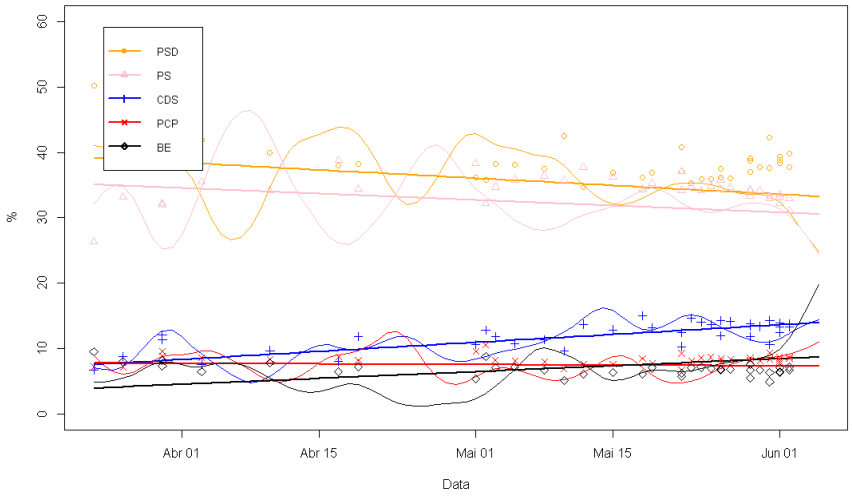


Figura 8.4: Interpolação da percentagem de *tweets* emitidos pelos media referenciando cada partido concorrente ás eleições legislativas durante a campanha. Linha de tendência de 1ª ordem minimizando o erro quadrático e valor das sondagem representados por símbolos pontuais. Total de 44 contas de media

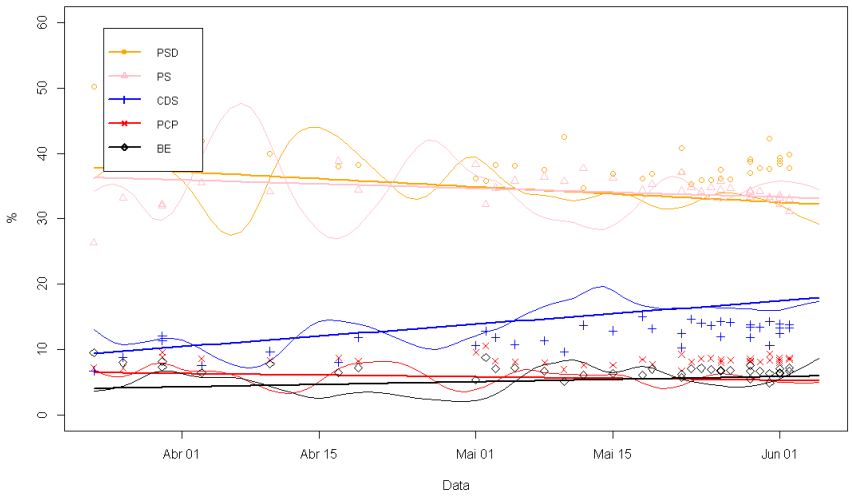
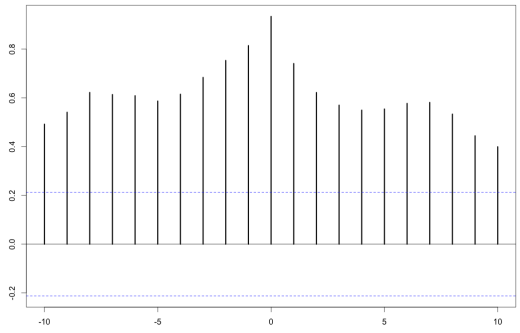
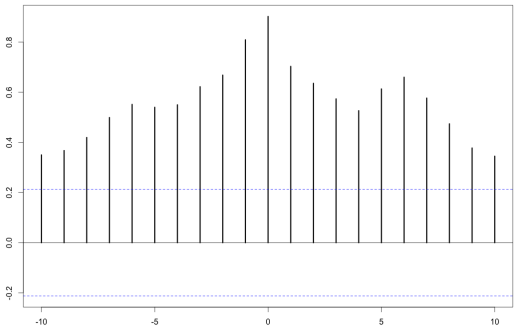


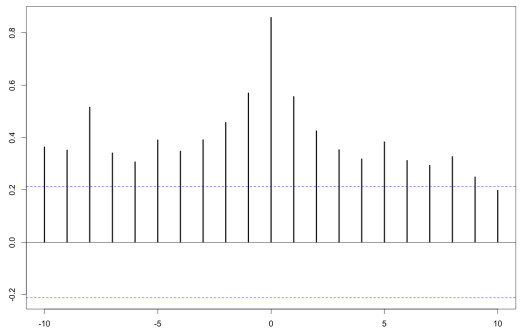
Figura 8.5: Interpolação da percentagem de *tweets* emitidos pelos utilizadores referenciando cada candidato presidencial durante a campanha. Linha de tendência de 1ª ordem minimizando o erro quadrático e valor das sondagem representados por símbolos pontuais. Total de 1903 utilizadores.



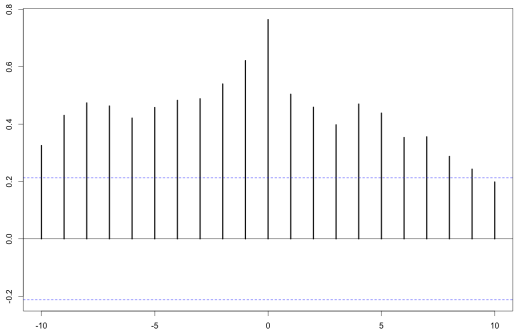
(a) Cavaco



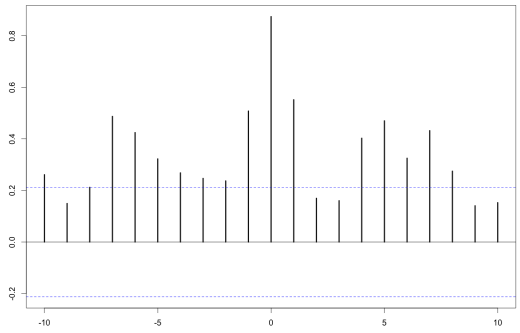
(b) Alegre



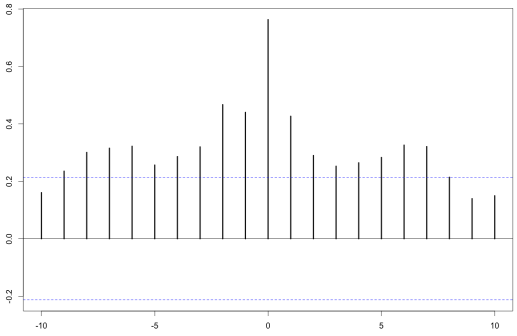
(c) Nobre



(d) Lopes

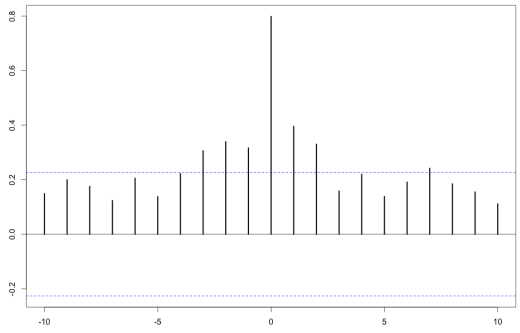


(e) Moura

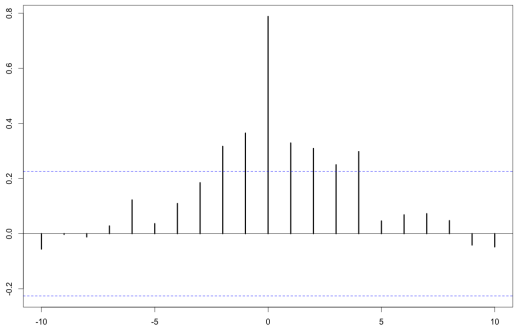


(f) Coelho

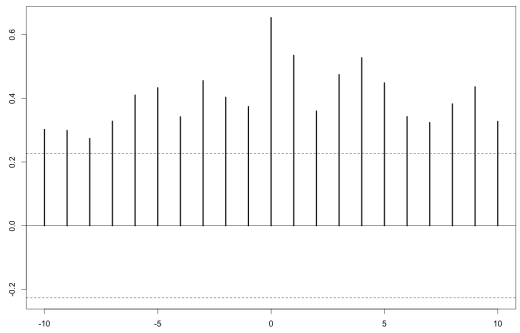
Figura 8.6: Covariance between time series of news and population tweets in presidential elections, lag of between -10 and 10 days.



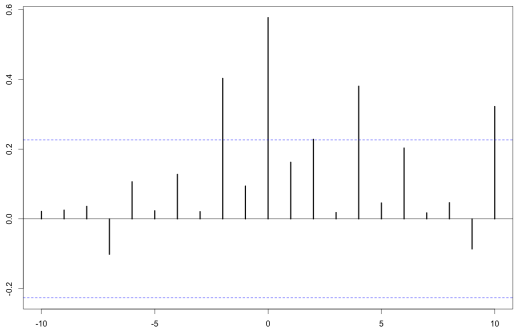
(a) PSD



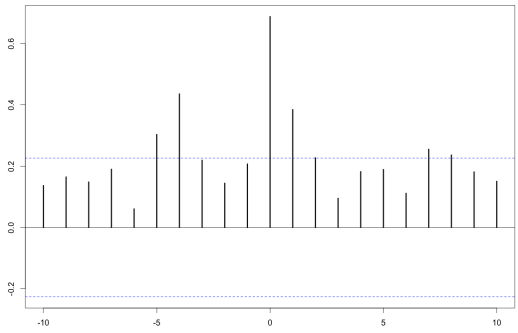
(b) PS



(c) CDS



(d) PCP



(e) BE

Figura 8.7: Covariance between time series of news and population tweets in legislative elections, lag of between -10 and 10 days.

Com o intuito de analisar este fenómenos em mais detalhe procurámos recolher outras intuições sobre o problem através de dois modelos sociofísicos que de seguida detalhamos.

8.3 Modelo browniano de influência

O modelo browniano de agentes de Frank Schweitzer para dinâmica de opiniões Schweitzer and Garcia [2010] que utilizamos foi concebido no seio de um enquadramento mais geral de modelo de agente social no qual cada agente é descrito por um conjunto de variáveis de estado $u_i^{(k)}$ onde o índice $i = 1, \ldots, N$ refere o agente individual i e k indica as diferentes variáveis. Estas variáveis podem ser *externas* e observáveis a partir do exterior do agente, ou podem constituir *graus de liberdade internos* que apenas se podem concluir das ações observáveis do mesmo agente Schweitzer [2007]. No caso geral a evolução dinâmica destas variáveis é modeladas segundo a equação diferencial:

$$\dot{u_i}^{(k)} = f_i^{(k)} + f_i^{estocastico} \tag{8.3}$$

Esta equação procura refletir o *principio da causalidade*: qualquer efeito, como a variação temporal de uma certa variável u tem algumas causas que são listadas no lado direito da equação. No conceito de agente Browniano é assumido que estas causas podem ser descritas pela sobreposição de influências determinísticas ($f_i^{(k)}$) e estocásticas ($f_i^{estocastico}$) no agente i. Esta distinção é baseada na ideia de Langevin da descrição do movimento Browniano das partículas que deu nome ao conceito Langevin [1908].

No caso dos agentes emocionais que exprimem opiniões e por elas são condicionados em comunidade, Schweitzer elaborou um modelo que obedece às seguintes variáveis externas:

$$\dot{a_i} = -\gamma_a a_i(t) + F_{a_i} + A_{ai}\xi_a(t) \tag{8.4}$$

$$\dot{v_i}^k = -\gamma_v v_i^k(t) + F_{v_i}^k + A_{vi}^k \xi_v^k(t) \tag{8.5}$$

Cada agente tem assim dois tipos variáveis de estado principais: *excitação* a_i, só com um modo, que caracteriza a predisposição dos agentes para sentirem as suas emoções, se expressarem ou agirem, e as *valências* v_i^k, com k modos, que representam valências emocionais arbitrárias do agente: *espanto*; *satisfação*; *alegria* ou quaisquer outras. No caso particular da nossa aplicação considerámos valências relacionadas com a simpatia ou aversão pelos candidatos ou partidos envolvidos nas eleições. Portanto $v_i^k \in \{'Cavaco', 'Alegre', 'Nobre', 'Lopes', 'Moura', 'Coelho'\}$ no caso das eleições presidenciais e $v_i^k \in \{'PSD', 'PS', 'CDS', 'PCP', 'BE'\}$ no caso das legislativas. Admitimos que a intenção do agente se expressar sobre um candidato ou um partido é determinada em simultâneo pelo seu nível de excitação pela valência que em si é mais saliente, de maior valor absoluto, no seio do seu conjunto interno de valências.

O primeiro termo do lado direito de ambas as equações está associado com o decaimento exponencial para o equilíbrio com as constantes temporais associadas γ_a and γ_v. O segundo termo F_{a_i} and $F_{v_i}^k$ reflete a característica resposta determinista dos agentes. O terceiro termo, ponderado por A_a and A_v^k respetivamente, ξ_a and ξ_v^k representa os fatores aleatórios envolvidos na resposta e que diferenciam a individualidade de cada agente.

Na implementação que fizemos do modelo seguimos de perto a configuração proposta por Garcia e Schweitzer Schweitzer and Garcia [2010] no entanto fizemos algumas alterações. Assim para termos um valor nulo no equilíbrio estável dos agentes na falta de informação externa, e de igual modo para obtermos soluções reais para a resposta da equação diferencial, escolhemos o termo F_v^k do seguinte modo:

$$F_v^k = h^k(t) v_i^k(t) \tag{8.6}$$

Neste caso a influência determinista é mais simples mas também mais explicativa do que a formulação original. O campo $h^k(t)$ representa um estímulo determinístico para a valência e está associado com a quantidade de notícias introduzida na comunidade relacionadas com essa mesma valência. Por exemplo $h^{Cavaco}(t)$ representa um estímulo para a valência de u_i^{Cavaco}. Deste modo

a valência é reforçada pela presença de notícias a ela associadas. Poderíamos ter escolhido outra função mais elaborada no entanto o poder de análise seria prejudicado pela introdução de processos de reforço de opinião de segunda e de terceira ordem como são referidos no artigo original.

Por seu lado a informação associada a cada valência deve ser moderada pelo debate na comunidade. Afim de modelarmos esta interação entre os agentes, condicionada pelo grau de excitação a_i, e a informação noticiosa, configurámos uma modulação contrária a $h^k(t)$ através de um campo $h(t)$ comum a toda a comunidade representando a discussão no coletivo dos agentes. Deste modo o cálculo de $h^k(t)$ é efetuado segundo a seguinte fórmula:

$$h^k(t) = n^k(1 - h(t)) \tag{8.7}$$

Onde n^k é a proporção de notícias acerca do candidato ou partido k e o campo $h(t)$ é obtido em função da excitação dos agentes:

$$\dot{h}(t) = -\gamma_h h(t) + N_a/N \tag{8.8}$$

Onde N_a representa o número de agentes tendo excitação $a_i(t)$ suficiente, ou seja quando $a_i > \tau$ para um valor de τ determinado. N é o número de agentes. Segundo esta formulação cada agente tem a noção da expressão dos media acerca de cada candidato ou partido mas também efetua uma interferência moderadora através do debate, que atenua o impacto das notícias. Em $h(t)$ seguimos também o modelo original de Schweitzer obedecendo a uma evolução dinâmica de decaimento exponencial γ_h, e a uma dependência inversa do número de agentes. Quando $a_i < \tau$ assumimos que o agente i não contribui para este campo.

A fórmula determinista para a excitação é dada por:

$$F_a = \hat{h}(t) \tag{8.9}$$

Onde $\hat{h}$ é a média dos $h^k(t)$ que influenciam cada valência v^k:

$$\hat{h}(t) = \langle h^k(t) \rangle \tag{8.10}$$

Ou seja, admite-se que o único fator determinístico que influencia o cresci-

mento da excitação dos agentes é o campo de notícias. Deste modo os agentes são excitados pelo número de notícias, mas uma excitação excessiva do coletivo atenua significativamente a sua valência específica.

Na Figura 8.9 encontram-se representado os fluxos de informação desta experiência.

O bloco *Tweets dos media* e *debate h(t)* representa como os agentes recebem vários sinais $h^k \in [0.0 \ldots 1.0[$ sinalizando a proporção de tweets da Comunicação Social que é misturada para cada candidato/partido com o campo de ruído de debate $h(t)$ que a comunidade produz (Equação 8.7).

O comportamento de $h(t)$ está dependente da constante de decaimento γ_h e também do número de agentes N_a com *excitação* acima de tau N_a(Equação 8.8).

Por ourto lado, os vários sinais h^k são injectados na componente determinística da *valência* F_v^k (Equação 8.6) e reforçados por ela, determinando por maioria o voto de cada agente (Equação 8.5).

Por outro lado, a dinâmica da *excitação* (Equação 8.4) depende também deste sinal através da sua componente deterministica F_a (Equação 8.9 e 8.10).

Por fim, no bloco losangonal, a valência global da comunidade, representada pela contabilização da votação de cada agente, e é que por sua vez determinada pela valência mais positiva de cada um, é comparada com o fluxo real de *tweets* que foi observado na comunidade. Este processo é efectuado através da medição da similaridade cosenoidal entre o vetor votação em cada instante dos dois fluxos C_r (Equação 8.11).

$$C_r = \langle \overrightarrow{\arg\max_k \{v_i^k(t)\}} \cdot \overrightarrow{T^k(t)} \rangle \tag{8.11}$$

O vetor $\overrightarrow{T^k}(t)$ representa o número de *tweets* coletados e tem 5 ou 6 dimensões no caso das eleições legislativas e presidenciais respetivamente.

Neste ponto convém fazer uma interpretação das constantes γ. Como constantes de amortecimento o seu efeito é o de atenuar a resposta impulsiva individual às influências determinísticas e estocásticas definidas na equação 8.3 do movimento browniano.

Na Figura 8.8 podemos comparar a magnitude seu efeito quando toma valores entre 0 e 1. Valores pequenos permitem uma resposta individual mais imediata

e sensível ás influências, valores maiores atenuam esses efeitos e provocam um efeito de *campo médio* entre individuos.

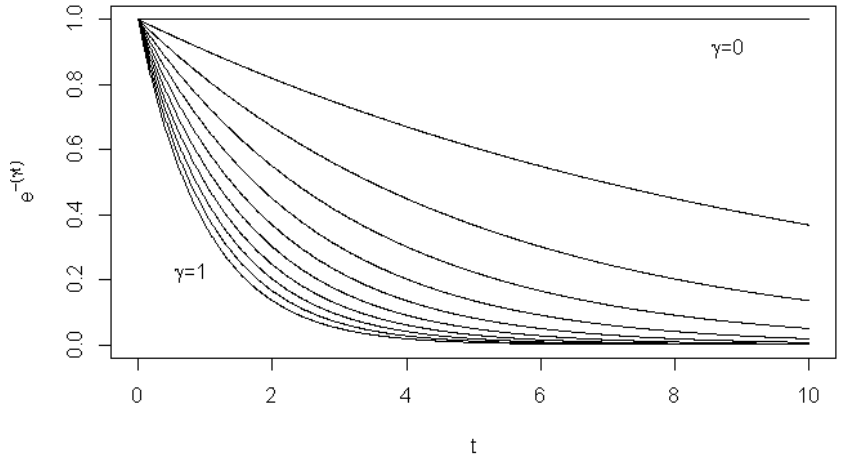


Figura 8.8: Decaimento exponencial para $\gamma \in [0 \ldots 1]$

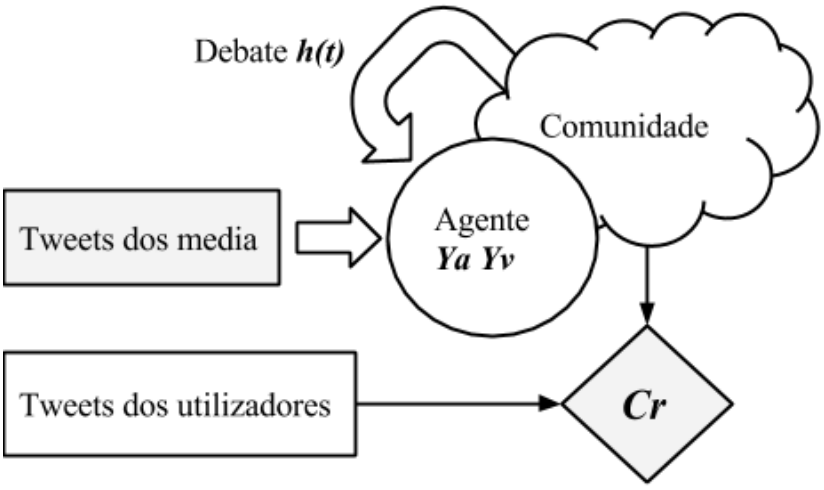


Figura 8.9: Diagrama da estrutura da simulação.

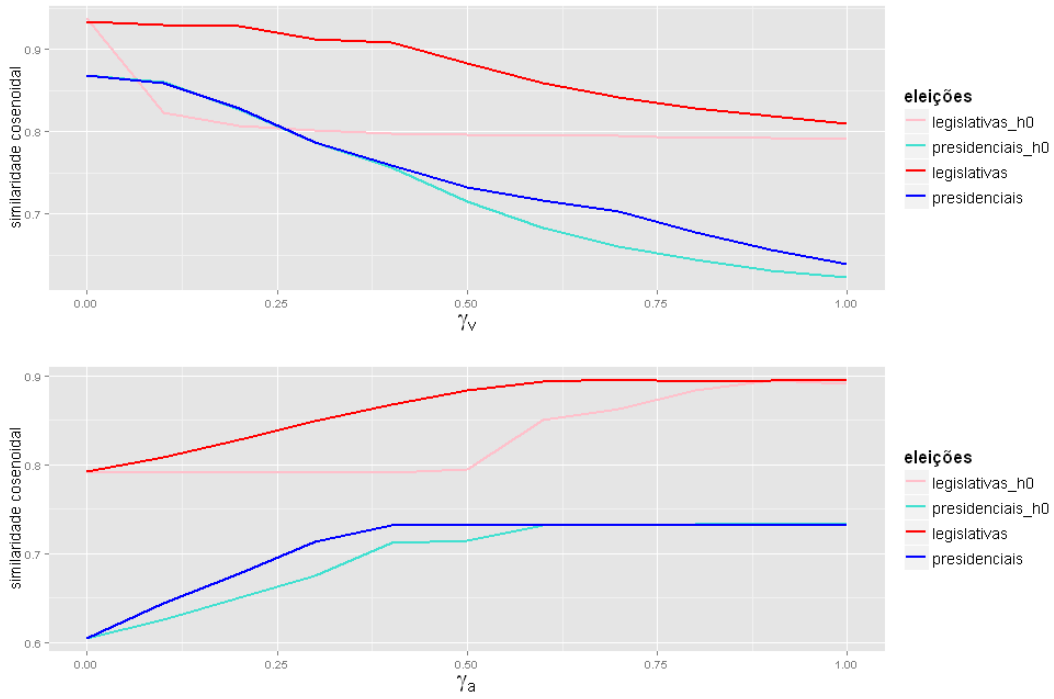


Figura 8.10: Resultados das simulações com uma comunidade de N=1000 agentes. Resultado médio de 20 corridas com $\gamma_a \in [0 \ldots 1]$, $\gamma_v \in [0 \ldots 1]$ e $\gamma_h \in [0 \ldots 1]$, os resultados não dependeram significativamente de γ_h exceto no caso particular em que $\gamma_h = 0$ que se encontra representado. O valor de $\tau = 0.5$.

8.3.1 Excitação no debate

Examinando a Figura 8.10 podemos verificar que a atenuação da dinâmica da respostas aos estímulos provocados pelas notícias (Equação 8.6), na valência dos agentes, prejudica a similitude da sua expressão com a expressão real dos *tweets* que foram recolhidos dos utilizadores. Ou seja, conforme pudemos observar nas Figuras 8.7 e 8.6 a resposta dos utilizadores em relação às notícias é relativamente imediata e a valência das expressões deve refletir de imediato a valência das notícias. Podemos também observar que esta resposta é bastante prejudicada quando o ruído do debate não é suavizado, $\gamma_h = 0$, em que pela Equação 8.8, $\dot{h}(t) = N_a/N$ e o campo informativo que reflete as noticias é permanentemente interferido pelo debate aleatório entre os agentes.

8.3.2 Valência do debate

Examinando a Figura 8.10 no respeitante à *excitação* confirmamos aquele facto. Assim se a excitação dos utilizadores não depender muito do nível de ruído noticioso e for atenuada através de um nível de γ_a elevado, a replicação dos resultados simulados é também mais próxima aos resultados recolhidos na vida real.

Desta experiência, apesar de se basear numa metáfora muito geral de possíveis dinâmicas de formação de opinião, podemos confirmar a grau e a velocidade de resposta dos agentes ás notícias pode condicionar a formação da opinião e concluir que:

- A resposta dos agentes à valência das notícias é rápida, ou seja, para os resultados de aproximarem dos reais é porque agentes respondem às notícias na hora em que são divulgadas. De facto uma análise superficial do conteúdo da nossa base de dados recolhida do *Twitter* confirma esse facto.

- O debate interno aceso contraria esse efeito. De igual modo podemos confirmar esses resultado. Não existe, na comunidade que pudemos acompanhar, um troca de opiniões prolongada. Os modos de comunicação referidos nas Figura 7.2 confirma esse fato. Grande parte da comunicação é feita de um para todos sem extensas réplicas. E quando é feita acompanha a valência das notícias.

Ambas estas conclusões partiram de pressupostos muito simples e de *campo médio*. Todos os agentes comunicam com todos e as notícias são por todos difundidas. Em seguida implementamos um segundo modelo mais elaborado em que os agentes são condicionados diretamente por outros agentes.

8.4 Modelo da teoria do impacto social

A partir da Teoria do Impacto Social, desenvolvida pelo sociólogo Bib Latané Latane [1981] em 1981, os sociólogos Andrzej Nowak a Jacek Szamrej desenvolveram nos anos 90 um modelo pioneiro de agente social que lhes permitiu observar fenómenos de polarização e de agregação de opiniões muito similares aos relatados por cientistas políticos em trabalhos nos Estados Unidos Nowak et al. [1990]. A Teoria do Impacto Social é muitos simples e baseia-se em três regras fundamentais:

- O impacto social é o resultado de forças sociais e depende da magnitude dessas forças
- A magnitude do impacto depende do número de fontes de impacto
- A magnitude do impacto tende a atenuar-se com o o aumento do número de alvos

Segundo a teoria este impacto provoca padrões dinâmicos de comportamento nos grupos com as seguintes propriedades:

- Consolidação - Com o decorrer das interações as opiniões uniformizam-se e a opinião da maioria tende a contaminar as opiniões das minorias.
- Agregação - Os indivíduos tendem a agregar-se em grupos de opiniões similares. Grupos de opinião podem emergir do todo e diferenciar-se.
- Correlação - Com o decorrer do tempo, as opiniões individuais tendem a correlacionar-se em diversos assuntos, mesmo nos que não são discutidos em grupo.

- Diversidade e continuidade - Pode existir diversidade num grupo, mas se a maioria é desproporcionada ou se os indivíduos em minoria não contactam entre si as diferenças esbatem-se.

Um modelo matemático possível para a teoria que implementámos na simulação multi-agente é representado no seguinte conjunto de equações Lyst et al. [2002]:

$$I_i = \sum_{j=1}^{N} \frac{p_j}{d_{ij}}(1 - \sigma_i\sigma_j) - \sum_{j=1}^{N} \frac{s_j}{d_{ij}}(1 + \sigma_i\sigma_j) \tag{8.12}$$

$$\sigma_i(t+1) = -sign(\sigma_i(t)I_i(t) + h_i(t)) \tag{8.13}$$

Na equação I_i é o impacto provocado pela comunidade no agente i . Cada agente i tem um certo grau de *persuasão* p_i e um outro grau de *suporte* s_i que refletem a força das interações sobre indivíduos com opiniões opostas ou com as mesmas opiniões σ_i respetivamente . A variável opinião - σ_i é bipolar e pode tomar os valores $+1$ e -1. A variável d_{ij} corresponde à distância entre os agentes, e a variável $h_i(t)$ representa ruído comunitário que interfere com o processo de formação de opinião.

Por forma a realizarmos o processo de votação, equação 8.13 foi ligeiramente alterada para:

$$\sigma_i^{k'}(t+1) = \sigma_i^k(t)(I_i^k(t) + h_i^k(t)) \tag{8.14}$$

com uma fase adicional de normalização:

$$\sigma_i^k(t+1) = \begin{cases} +1 & k = \arg\max_k(\sigma_i^{k'}(t+1)) \\ -1 & \text{outros casos} \end{cases} \tag{8.15}$$

Esta normalização substitui a operação $-sign()$ em 8.13. A variável h_i^k em 8.14 representa o fluxo de informação que entra na comunidade. A modelação do processo eleitoral consiste na adaptação do modelo original de opinião tendo cada agente um vetor de opiniões σ_i^k com $k \in \{'Cavaco', 'Alegre', 'Nobre', 'Lopes', 'Moura', 'Coelho'\}$ or $k \in \{'PSD', 'PS', 'CDS', 'PCP', 'BE'\}$.

Tendo então em consideração a rede social de ligações entre os utilizadores procurámos analisar o impacto desta estrutura na formação da magnitude relativa de expressões nos dados que recolhemos. Antes de considerarmos este problema vamos contextualizar a questão da estrutura analisando a rede dos utilizadores.

8.4.1 A estrutura da rede

Na rede social *online* Twitter os utilizadores estão agrupados por relações de seguimento e amizade que permitem a troca de informação. Tem sido argumentado que as redes sociais apresentam uma topologia com distribuição de graus aproximada por leis de potências. Alguns estudos afirmam que as redes de Barabasi-Albert criadas por processos mistos de *conexão preferencial* e *crescimento uniforme* apresentam um comportamento em lei de potências Albert and Barabási [2002]. De facto os processos de *conexão preferencial* estão presentes em muitas redes sociais, naturais e tecnológicas. Desde as redes de *routers* na Internet, de atores de Hollywood, citações em artigos científicos, tamanhos de cidades e outros fenómenos o mecanismo de *conexão preferencial* parece explicar bem a génese da topologia e o processo de crescimento Newman [2010]. No entanto, tendo em conta a distribuição de graus de conexão na rede que recolhemos, presente na Figura 8.11, verificamos que existe uma discrepância acentuada com uma distribuição simples em lei de potências. Recentemente outros autores adiantaram uma explicação generativa para as redes sociais *online* como a rede Twitter mais ajustadas à sua topologia. Esta explicação suporta-se num processo misto de lei de potências e de distribuição log-normal que os autores nomearam como distribuição Pareto-Lognormal Sala et al. [2011]:

$$f(x) = \beta x^{\beta-1} e^{(-\beta\mu+\frac{\beta^2\tau^2}{2})} \Phi^c(\frac{\log x - \mu + \beta\tau^2}{\tau}) \tag{8.16}$$

Utilizando os valores $\beta = 1.2$; $\mu = 4.0$ e $\tau = 1$, confirmamos também este ajustamento neste caso de estudo que pode indiciar um processo generativo formulado na equação 8.16. Os autores sugerem um algoritmo em dois passos que integra propriedades fundamentais da *lei dos efeitos proporcionais* e da *conexão preferencial*. O algoritmo alterna entre a adição de novos nós à rede segundo um modelo de *conexão preferencial*, e crescer essa conectividade entre nós pela *lei*

dos efeitos proporcionais. De facto estes dois processos têm respaldo na lógica de conexão de uma rede social como é a rede de amizade e de troca de informação do Twitter:

- Os utilizadores populares tendem a atrair mais conexões, mas também tendem a conectar-se eles próprios mais com outros utilizadores menos conectados.

Este processo é particularmente visível no Twitter uma vez que as ligações originadas pelo utilizador e as que lhe são afetas por outros são discriminadas na própria rede. Estes processos também são conciliáveis com outros estudos sobre *capital social* (o número das suas ligações) *capital semântico* (diversidade dos conteúdos) e a assortatividade, em redes sociais blogues Roth and Cointet [2010], onde é referida uma ligeira disassortividade entre utilizadores uma fraca correlação entre os dois capitais, favorecendo portanto a diversidade no tipo das relações.

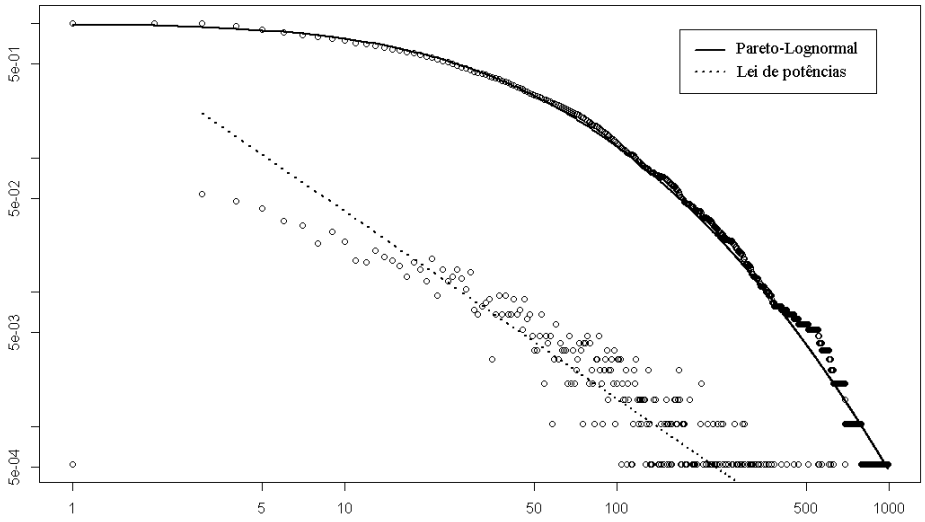


Figura 8.11: Distribuição de grau na rede de utilizadores recolhida e distribuição complementar cumulativa da mesma rede. Comparação com um ajustamento em lei de potências ($\alpha = 1.396$) e com um ajustamento à função cumulativa por função Pareto-Lognormal.

8.4.2 A topologia da rede e a expressão dos agentes

Com o propósito de analisarmos o impacto da topologia no debate entre os agentes, efetuamos simulações modelizadas segundo o esquema da Figura 8.9 com o modelo de influência por impacto social atrás descrito. A parametrização do modelo foi efetuada tentando eliminar o número de graus de liberdade. Aos parâmetros s_i e p_i foi atribuída uma distribuição gaussiana, com média 1.0 e variância 0.5 e ao parâmetro d_{ij} foi atribuído a valor unitário e correspondente à ligação entre os vizinhos na rede real. Deste modo procurou-se estudar o fenómeno de influência em paralelo com o caso real.

Na primeira fase simulámos a comunidade de agentes modificando parcialmente a topologia da rede original por forma a detetarmos a significância desse alteração na propagação da influência das notícias. Nas Figuras 8.13 e 8.12 está figurado o erro total entre a expressão sobre candidatos e partidos recolhida na

rede social e a expressão contabilizada na simulação em função da progressiva reconexão aleatória das ligações.
Apesar de no caso das eleições presidenciais a introdução da aleatoriedade não incrementar muito o erro, no caso das eleições legislativas a deterioração do erro com a modificação aleatória das ligações é notória.

É conhecido que a reconexão aleatória tende a diminuir o caminho médio entre quaisquer dois nós da rede Newman [2010] transformando a redes regulares em mundos pequenos (*small world network*) Watts [2004]. Com a diminuição do caminho médio e a aproximação entre os agentes o impacto social e a influência entre agentes aumenta na comunidade. Podemos assim confirmar neste caso o que já tinha sido verificado na simulação anterior com o modelo Browniano: o aumento do debate e da contaminação de opiniões entre agentes implica uma redução da influência dos media, pois os agentes, conforme reportado nas figuras 7.2 e 7.3, tendem a expressar-se de si para fora e não entre eles. Como no caso anterior o comportamento nas presidenciais em função da notícia é pior.

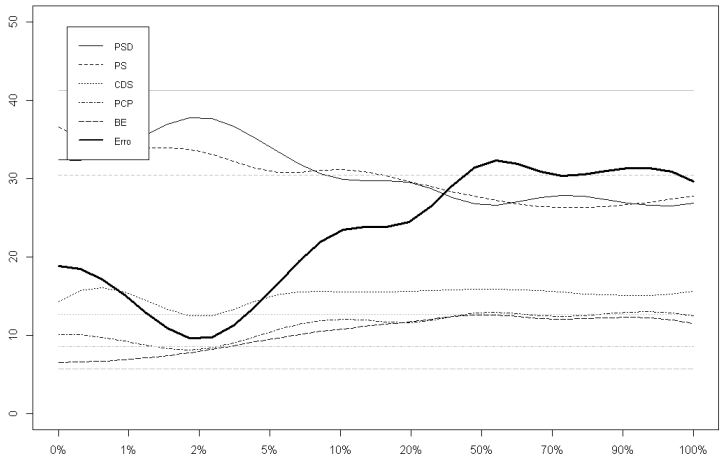


Figura 8.12: Gráfico do erro em valor absoluto de percentagem (linha escura) entre a estimativa de máxima verosimilhança da média de 30 corridas da simulação e o resultado final das eleições em função da percentagem de reconexão aleatória da rede. Eleições legislativas. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.527, 0.487, 0.365, 0.405, 0.372, 0.325. N=1903 agentes, cobertura mediática a 60% dos agentes.

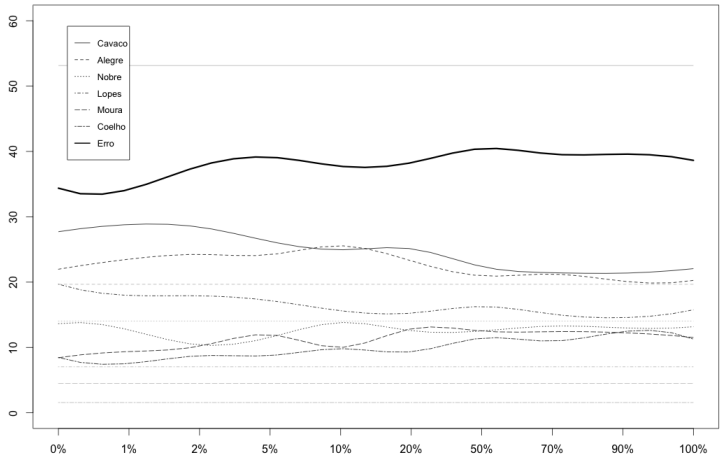


Figura 8.13: Gráfico do erro em valor absoluto de percentagem (linha escura) entre a estimativa de máxima verosimilhança da média de 30 corridas da simulação e o resultado final das eleições em função da percentagem de reconexão aleatória da rede. Eleições presidenciais. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.563, 0.529, 0.395, 0.343, 0.265 . N=1903 agentes, cobertura mediática a 60% dos agentes.

8.4.3 O impacto da abrangência noticiosa

Com o propósito de confirmarmos o grau de influência das notícias na resposta dos agentes, e de analisarmos o papel da topologia particular da rede nessa influência externa efetuámos dois conjuntos de simulações com duas redes distintas:

1. a rede original - com a distribuição de grau reportada na Figura 8.11.
2. uma malha regular de geometria retangular - todos os agentes têm 4 vizinhos contíguos.

Neste caso variámos a cobertura das notícias representada na equação 8.14 entre 10% e 100% dos agentes de modo aleatório. Em cada ensaio é escolhida uma percentagem de agentes que recebe noticias sendo os restantes meramente influenciados. Deste modo procurámos avaliar o impacto da cobertura na repetição dos resultados reais.

Conforme podemos observar no conjunto de Figuras 8.14; 8.15; 8.16 e 8.17, a progressiva cobertura de notícias diminui o erro entre os resultados das simulações e os resultados recolhidos na rede *Twitter* original. Este facto é, como já verificámos, muito mais acentuado no caso das eleições legislativas.

No entanto outra observação podemos fazer: no caso da malha regular, onde portanto os agentes não são diferenciados, este decréscimo do erro é muito menos pronunciado. Para além disso o desvio entre os agentes, a variância da sua resposta é muito menor, como podemos comprovar pelos erros do estimador na legenda.

Não havendo outro fator diferenciador entre os agentes que não o grau das ligações, uma vez que a diversidade de p_i e s_i é anulada na repetição dos ensaios, podemos concluir que há uma relação entre a diversidade deste grau e a emulação dos resultados reais. De algum modo a diversidade do grau das relações, do capital social, mantém os indivíduos menos dependentes da influência entre si, não deixando de permitir a difusão de informação. É este o principio básico subjacente aos conceitos de *brokerage* e *closure* Burt [2005], introduzidos em 2005 por Ronald Burt. O isolamento entre os indivíduos através de mediadores permite a diversidade de opiniões consensuais sem eliminar a possibilidade de difusão da informação e da difusão da inovação.

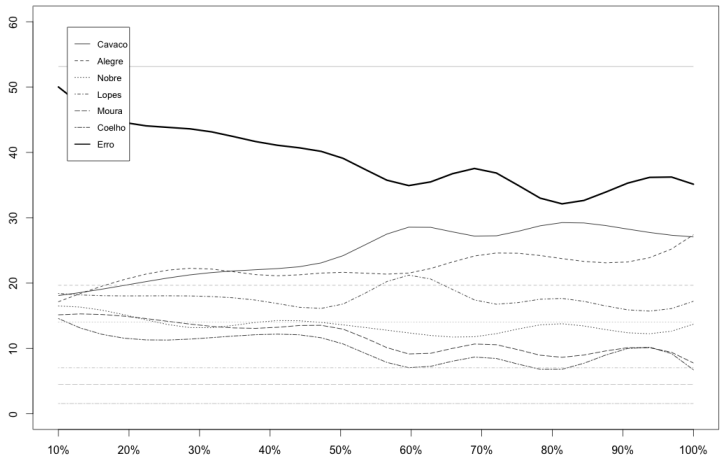


Figura 8.14: Gráfico do erro em valor absoluto de percentagem da diferença (linha escura) entre um estimador de máxima verosimilhança da media de 30 corridas da simulação e o resultado final das eleições, função da percentagem de cobertura de notícias pela rede de agentes. Rede original. Eleições presidenciais. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.464, 0.417, 0.328, 0.430, 0.301, 0.267. N=1903 agentes.

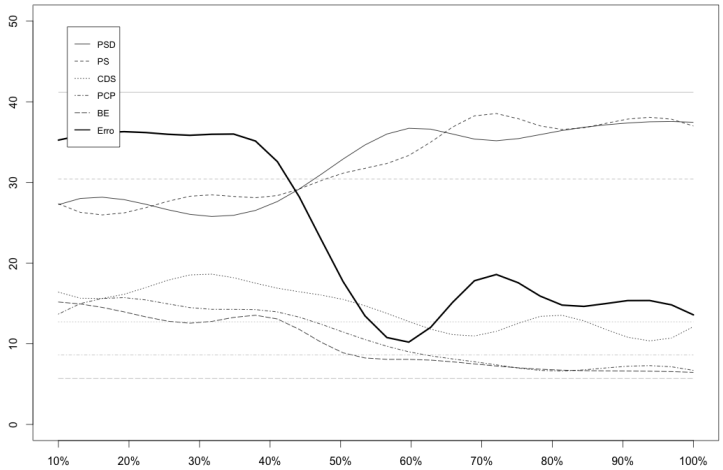


Figura 8.15: Gráfico do erro em valor absoluto de percentagem da diferença (linha escura) entre um estimador de máxima verosimilhança da media de 30 corridas da simulação e o resultado final das eleições, função da percentagem de cobertura de notícias pela rede de agentes. Rede original. Eleições legislativas. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.501, 0.472, 0.378, 0.305, 0.246. N=1903 agentes.

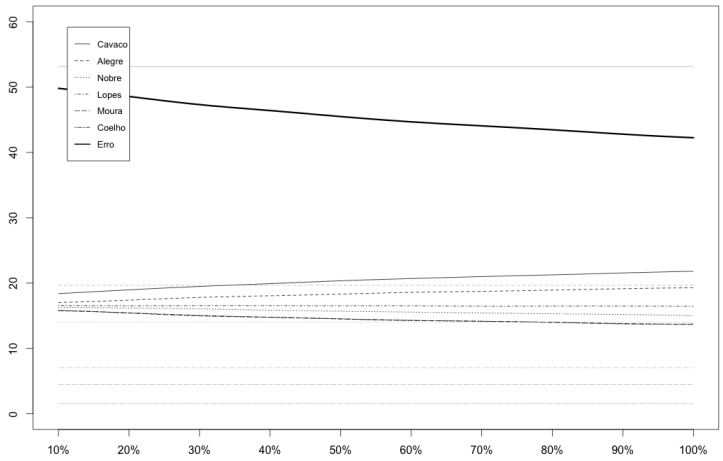


Figura 8.16: Gráfico do erro em valor absoluto de percentagem da diferença (linha escura) entre um estimador de máxima verosimilhança da media de 30 corridas da simulação e o resultado final das eleições, função da percentagem de cobertura de notícias pela rede de agentes. Malha regular. Eleições presidenciais. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.173, 0.057, 0.050, 0.053, 0.048, 0.044. N=1903 agentes.

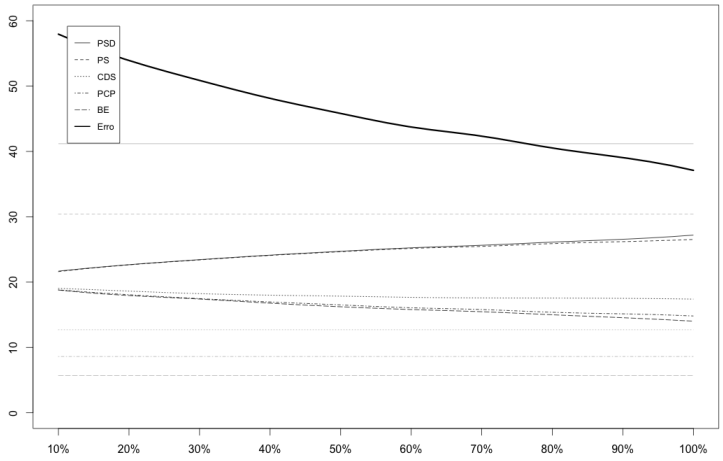


Figura 8.17: Gráfico do erro em valor absoluto de percentagem da diferença (linha escura) entre um estimador de máxima verosimilhança da media de 30 corridas da simulação e o resultado final das eleições, função da percentagem de cobertura de notícias pela rede de agentes. Malha regular. Eleições legislativas. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.121, 0.090, 0.083, 0.065, 0.070. N=1903 agentes.

8.4.4 O efeito da memória na teoria do impacto social

Em último lugar efetuámos um conjunto de experiências afim de medirmos a influência de atrasos nos impactos relativamente ao fluxo das notícias que influenciam a comunidade. Contrariamente ao modelo browniano, o modelo clássico de impacto social não contempla nenhum efeito de memória, o estado dos agentes no tempo t não depende do seu estado no tempo $t-1$. Por forma a incluirmos este efeito nos ensaios introduzimos uma variável δ que provoca um atraso no impacto que os agentes provocam entre si. A equação 8.14 foi reescrita da seguinte forma:

$$\sigma_i^{k'}(t+1) = \sigma_i^k(t)[I_i^k(t) + \delta I_i^k(t-1) + h_i^k(t)] \tag{8.17}$$

As figuras 8.18 and 8.19 reportam os efeitos que esta variável induz nas simulações. Pode-se observar, como acontece nos outros gráficos, a diferença entre as eleições presidenciais e as legislativas onde o erro da expressão simulada dos agentes relativamente à expressão real é menor. De novo confirmamos o que já tínhamos verificado nas figuras 8.6 e 8.7: a correlação muito forte entre a notícias e os *tweets* dos utilizadores e o facto da resposta destes ser efetuada em simultâneo com as noticias e sem um atraso significativo.

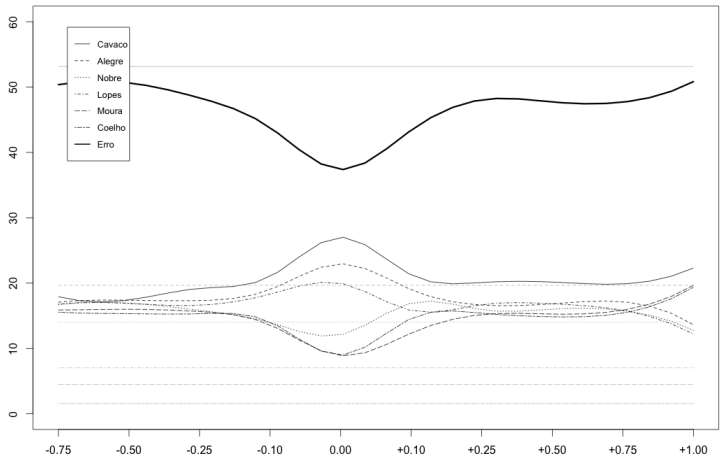


Figura 8.18: Gráfico do erro em valor absoluto de percentagem da diferença (linha escura) entre um estimador de máxima verosimilhança da media de 30 corridas da simulação e o resultado final das eleições, função do atraso δ entre o estimulo das notícias e o impacto. Rede original. Eleições presidenciais. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.442, 0.416, 0.302, 0.388, 0.248, 0.245. N=1903 agentes, cobertura noticiosa de 60% dos agentes.

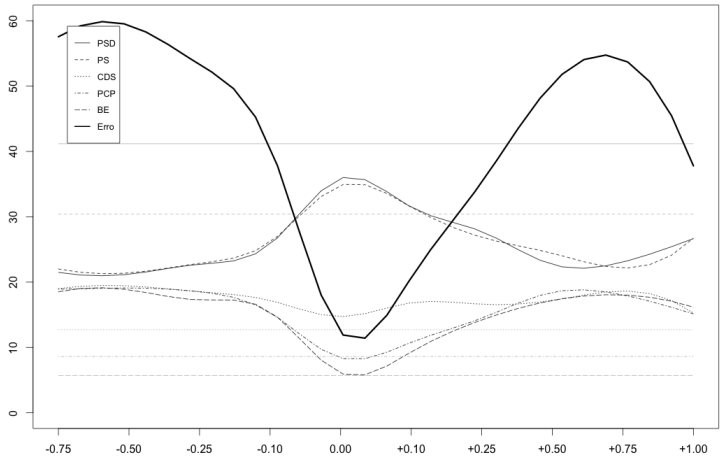


Figura 8.19: Gráfico do erro em valor absoluto de percentagem da diferença (linha escura) entre um estimador de máxima verosimilhança da media de 30 corridas da simulação e o resultado final das eleições, função do atraso δ entre o estimulo das notícias e o impacto. Rede original. Eleições legislativas. A cinzento a votação simulada dos agentes. Erro máximo da média entre corridas : 0.520, 0.518, 0.432, 0.352, 0.323. N=1903 agentes, cobertura noticiosa de 60% dos agentes.

8.5 Conclusões do estudo de caso

Em resumo listamos de seguida as principais conclusões que retiramos deste nosso caso de estudo:

1. Os fluxos de *tweets* durante os três meses que antecederam as eleições apresentam uma magnitude relativa que segue de perto a quantidade de notícias produzidas sobre as mesmas entidades. No *Twitter* em particular as pessoas tendem a falar sobre as notícias Kwak et al. [2010] e acerca das notícias do dia. Verificámos este facto não só a partir do conteúdo dos *tweets* mas também pela validação desta hipótese através de simulações. Mais relevante foi o facto de descobrirmos que a proporção de *tweets* de agentes noticiosos ou *tweets* de utilizadores, representar uma boa estimativa quer do valor corrente das sondagens eleitorais clássicas quer do resultado final das eleições. Com menos precisão do que as sondagens clássicas, o resultado eleitoral relativo, quando empates estreitos não estão em jogo, pode ser previsto.

2. Num ensaio de simulação do debate no seio da comunidade sem ter em conta a topologia das ligações, mas apenas um efeito de campo médio da excitação dos agentes e da sua valência relativamente aos candidatos e partidos, utilizamos um modelo browniano de agente. Verificámos que para obtermos resultados similares aos reais os níveis de flutuação de excitação dos agentes deveriam ser atenuados. Contrariamente, os níveis de flutuação da valência deveriam ser deixados livres. Estes resultados em termos práticos significam que os agentes se devem manifestar sem muita interação, uma vez que esta impacta negativamente os resultados, mas com relativa liberdade de reagir ao estimulo das notícias. De novo estes resultados são confirmados pelo estudo do conteúdo observado nos *tweets*. Os utilizadores tendem a debater pouco e a expressar-se para todos, com relativo imediatismo, a partir dos conteúdos noticiosos que absorvem.

3. A partir da análise da distribuição de grau da rede de utilizadores que coletamos, verificamos que a mesma é coincidente com uma distribuição

Pareto-Lognormal, cujo modelo generativo tem diversos paralelos com outros estudos sobre redes sociais *online*, nomeadamente a dissasortividade nas relações. Os nós com mais capital social tanto se ligam a nós mais isolados como entre si.

4. Finalmente efetuámos um ensaio de simulação tendo em conta a topologia das ligações e não uma influência a partir de um campo comum. Para este estudo utilizámos um modelo baseado na Teoria do Impacto Social. Este modelo mostrou que a topologia desempenha um papel importante na obtenção de resultados similares aos reais, pois um ensaio com uma topologia radicalmente diferente, regular e em malha, mostrou uma degradação muito forte. De igual modo verificámos que a cobertura mediática dos agentes tem um papel importante, mas que também tem um papel limiar, com um ponto de viragem, a partir do qual não incrementa muito os seus efeitos. Finalmente verificámos também, confirmando os resultados que já tínhamos obtido, que existe uma coincidência temporal forte entre os *tweets* das noticias e os *tweets* dos utilizadores, pois o impacto atrasado dos primeiros degrada bastante a obtenção de resultados similares aos reais.

Parte V

Conclusão

Capítulo 9

Discussão e Perspectivas

Este capítulo conclui o documento de tese com a discussão das propostas e dos resultados apresentados e com algumas perspectivas de investigação dos trabalhos de tese.

9.1 Discussão

9.1.1 A popularidade das entidades a longo prazo obedece a padrões de distribuição probabilística

O primeiro resultado relevante nesta tese diz respeito à conclusão sobre um dos mecanismos de formação da popularidade.

O resultado que obtivemos diz-nos que o crescimento a longo termo da popularidade de uma determinada entidade, fruto de processos de comunicação entre individuos, é independente das qualidades subjectivas daquela ou destes, dependendo do processo em que é comunicada.

Este processo, conforme a formalização que fizemos, pode suportar-se num conceito de *mensagem* equivalente à noção de *dados*, uma entidade que veicula informação, que no entanto não a determina ou quantifica e desprovida de um valor de veracidade. Este tipo de definição permite avaliar o mecanismo de génese de maneira quantitativa, influenciado apenas na forma como esses dados transitam entre individuos.

Este processo obedece a um mecanismo semelhante ao da *Lei dos Efeitos Proporcionais* de Gibrat, cujo taxa de crescimento pode ser representada por uma variável aleatória, correspondente às imponderâncias de cada acto de replicação da mensagem que veicula a popularidade mas que obedece a um efeito multiplicativo, entidades mais populares crescem em popularidade mais rapidamente, de tal modo que o crescimento agregado obedece não a uma lei Normal ou Gausssiana mas Lognormal. O crescimento é no entanto modulado pelo efeito de atenção à novidade conforme explicitado na equação 4.25.

Os dados experimentais que recolhemos parecem confirmar esta tese, pois adequam-se melhor a este tipo de distribuição, incluindo os factores de estiramento da curva Lognormal associados à novidade.

Um sério contendor desta formalização é presente no estudo de Rajeev Kohli and Raaj Sah Kohli and Sah [2004], no qual através do clássico modelo Dirichlet-Multinomial de Goodhardt et al Goodhardt et al. [1984] concluem que a distribuição de vendas, portanto da popularidade de produtos, obedece a uma distribuição em Lei de Potências numa larga classe de produtos e marcas. Por um lado, dada a similitude das duas distribuições na cauda, poderá haver, uma vez que não foi medido, um melhor ajuste ao nosso modelo nos dados avançados pelos autores. Por outro lado, o modelo dos autores baseia-se numa distribuição prior,a Dirichlet, onde existe rivalidade nas escolhas entre marcas que está suposta no próprio conceito de quota de mercado (A distribuição de preferências, probabilidade de aquisição p_i é Dirichlet com $\sum_i p_i = 1$). O nosso modelo contempla uma oferta dinâmica de entidades(marcas) e novidade, mais contextualizado nos dados que nos serviram de teste.

Os modelos mais recentes de crescimento de firmas, onde os factores de crescimento são mais difusos e destes alguns não rivais, comprovam melhor os nosso resultados Sutton [1997] Growiec et al. [2008].

9.1.2 A dinâmica de evolução da popularidade segue um função simples

Um segundo resultado desta tese, comprovado pelos dois modelos dinâmicos de crescimento da popularidade apresentados, indica que a taxa de crescimento da popularidade pode ser aproximada por equações simples. Esta conclusão supõe, no entanto, que a popularidade é formada num modelos multiplicativo como o anterior, no nosso caso um modelo de ramificação ou um modelo epidémico.

Obtivemos perfis típicos de evolução temporal de popularidade de *memes* em blogues e *hashtags* na rede social Twitter a partir de amostras significativas. Estes perfir serviram para validar os modelos. Verificou-se que o modelo de ramificação, em que a resposta a um evento singular, exterior à comunidade e abrangendo de uma vez uma grande porção desta, seguida de uma propagação multiplicativa da notícia com um decaímento exponencial no tempo explica bem os resultados experimentais. Quando neste modelo a memória do eventos passados é representada no processo de sucessiva replicação de mensagens segundo a função $e^{-\theta t}$ os perfis experimentais ajustam-se muito bem. Por outro lado, o modelo epidémico explica, a partir de uma iniciação singular no interior da comunidade, os perfis experimentais quando a taxa de infecção epidémica é dada pela equação 6.19. As epidemias difusoras de popularidade são sobreponíveis. Os ajustamentos que fizemos segundo este método mostraram corresponder a processos paralelos que quando iniciados a tempos distintos se ajustam bem aos ciclos diários dos perfis, quando existem.

Verificámos também que os dois modelos se podem sobrepôr. A probabilidade das mensagens serem escutadas pode ser determinada pela topologia das relações intrasociais onde a mensagem se difunde, a probabilidade de ela ser replicada, pode estar associada ao valor da expressão da mesma para os individuos. Em conjunto, ambas determinam o crescimento da popularidade. Por outro lado, a probabilidade da mensagem ser esquecida, também ligada com o facto de não ser replicada, determina a velocidade com que as epidemias ou ramificações se dissipam. Neste contexto, pode ser definida uma equivalência, por aproximação, entre os dois modelos conforme está explicitado na equação 6.22.

9.1.3 A popularidade dos candidatos nas redes sociais é indicador dos resultados eleitorais

Um dos resultados mais significativos do ponto de vista empírico que obtivemos é o que o número de menções aos candidatos, ou aos partidos, na rede Twitter, em tempo de campanha eleitoral, segue de perto as percentagens das sondagens e reflete-se no voto final. Este resultado parece indicar que há uma forte correlação entre a expressão das pessoas e a sua intenção de voto. Ou seja, pelo menos em termos eleitorais, a magnitude da expressão é um forte indicador de preferência e a popularidade parece induzir essa preferência.

9.1.4 O efeito da comunicação social é determinante para a popularidade

Outro resultado empírico importante registado mostrou que existe uma influência forte entre a expressão da Comunicação Social e a expressão dos indivíduos. Este resultado foi testado de diferentes modos.

Foi examinada a correlação temporal entre os fluxos de mensagens sobre os candidatos/partido emitidos pela Comunicacão Social e pelos individuos. Verificou-se que existe uma forte correlação temporal, muito imediata, entre os dois fluxos e que esta correlação pode constituir uma medida da influência dos media na opinião pública.

Seguindo a mesma lógica foram posteriormente testados modelos de influência social para ser examinada a dinâmica da opinião entre as diversas valências políticas.

No primeiro modelo, um modelo browniano em que os agentes se comportam como partículas agitadas por campos de influência, examinámos a expressão aleatória da valência dos agentes influenciada quer pelo campo dos media quer pelo debate interno. Tratando-se de um modelo de campo médio a topologia das relações particulares dos agentes não é avaliada. No entanto foi possível verificar que, em simulação, a resposta imediata dos agentes ás notícias se coaduna mais com uma correlação próxima aos resultados reais.

De igual modo a agitação dos agentes deteriora a correlação com os resultados reais. A influência direta dos media sobre os agentes é deteriorada se estes interagem muito. Este último resultado é natural de esperar onde existe interação global de todos com todos. Devido a este facto, testámos o mesmo modelo de interação, desta vez com um modelo de influência diferente baseado num modelo sociológico inspirado na Teoria do Impacto Social, em que a topologia das relações é importante. Neste modelo verificámos que a topologia natural da rede favorece a expressão dos agentes, ou seja, quando a rede é aleatóriamente reconectada, aproximando-se de uma rede em *mundo pequeno*, os resultados afastam-se dos reais. Este efeito não é trivial, uma vez que é obtido por simulação. No caso limite de uma rede em malha regular a influência entre individuos é de tal forma uniforme que a expressão global das diversas valências se degrada.

A diversidade das expressões tem a ganhar com a diversidade dos modos de relações.

Confirmando os resultados anteriores, verificámos que a maior cobertura noticiosa da comunidade influencia a similitude com os resultados reais. Apesar disso a partir de um limiar de 60% de cobertura os resultados não progridem mais. Concluímos que que o resto da propagação da influência é deixada à própria comunidade.

Também foi testado o efeito do atraso entre o estímulo noticioso e o impacto entre os agentes que veio reconfirmar a imediatez destes na resposta ás notícias.

9.2 Perspectivas

O trabalho de investigação que agora reportamos representa apenas uma contribuição para o estudo dos fenómenos da comunicação na sociedade. Na mesma linha quantitativa que empreendemos, de analisar a difusão de informação, diversos campos estão já abertos e muitos se irão abrir à exploração e à investigação.

Apesar de não termos aprofundado o tema, nomeadamente no modelo de ramificação, a influência da geometria das relações sociais é um trabalho que pode ser aperfeiçoado neste modelo. De facto, o modelo contempla um termo que depende da densidade das conexões entre indivíduos por onde se propaga a informação.

É conhecida da teoria do capital social Burt [2005] a importância do papel dos mediadores (*brokers*) de informação na difusão desta entre comunidades. Estes indivíduos, alguns como fazedores de opinião, desempenham um papel relevante ao facilitar a passagem da palavra entre comunidades ao invés separadas. De igual modo podem ser centrais na formação de popularidade. Qualquer meio de comunicação social pode ser considerado como tal, se o entendermos como *brokers* da informação entre comunidades. É portando da mais fundamental relevância o estudo da topologia das relações nesta área.

Relativamente à predição muito há a fazer no sentido de aproveitar os dados sobre a sociedade que indiretamente são fornecidos pelas plataformas na Internet. Grandes empresas como a Google já o estão a empreender [1]. A função preditiva de modelos como o que propusemos nesta tese, pode não só ter resultados lucrativos bem como desempenhar um papel útil e orientador na gestão pública, quer dos espaços mediáticos quer de outros recursos comuns.

De igual modo o papel desempenhado pela memória não foi presente neste trabalho. O estudo da popularidade associado à função memorial, à repetição e insistência e ao calendário não foi abordado. Estes fatores são com certeza relevantes, nomeadamente na formação da memória coletiva e no papel desempenhado pela popularidade na formação da cultura.

[1] Ver por exemplo o caso do Google Domestic Trends da Google Finance http://www.google.com/finance/domestic_trends (acedido em Outubro de 2014) ou do trabalho efetuado nos laboratórios das grandes empresas de redes sociais online como o Facebook e o Twitter

Apêndice A

Listagem de um modelo NetLogo que exemplifica a distribuição de mensagens numa comunidades de agentes.

Cada agente possui uma memória (ou consciência) chamada `token` cujo tamanho é igual para todos e pré-determinado antes de se iniciar a simulação. Existem também neste mundo um conjunto de partida de N tokens distintos. O valor de N também é determinado antes da simulação. Cada um destes N tokens possui um determinado grau de *fitness* que pode ser igual para todos ou aleatório com uma distribuição Gaussiana de desvio padrão 0.2 e média que pode ser definida entre 0 e 1. Para além destes parâmetros existe uma probabilidade de `novidade` que determina se em cada iteração um token é adicionado à memória do agente.

A simulação é muito simples, em cada iteração, para cada agente:

1. Com uma probabilidade `novidade` um token é adicionado à memória do agente.
2. Se um token escolhido ao acaso na memória de um dos outros agentes tiver um fitness maior que um número aleatório com distribuição uniforme entre 0 e 1 ele passa a fazer também parte da memória do agente.
3. A variável `token` do agente é baralhada.
4. São descartados os tokens na variável `token` que excedam o tamanho máximo da memória.

5. O agente apresenta a cor do primeiro token que tiver na sua memória (ou consciência).

Este pequeno modelo simula os efeitos da variável γ_i através do *fitness* de cada token quando estes são trocados entre agentes.

Após algumas iterações a popularidade dos token com mais fitness sobressai. No entanto, como referimos no texto, se todos os tokens tiverem o mesmo fitness, a popularidade de um deles sobressai acabando por popular globalmente a memória (consciência) de todos.

Os objectos da interface são os seguintes:

- Slider `num_agentes` 0 a 1000
- Slider `fitness` 0 1.0
- Slider `num_tokens` 0 a 2000
- Slider `novidade` 0 1.0
- Slider `tam_memoria` 0 100
- Interruptor `aleaorio`
- Botão `inicializa` corre `inicia`
- Botão `go` corre `go` (fixo)
- Botão `reset` corre `reset`
- Gráfico `Numero de Tokens` desenha o número de tokens diferentes mostrados pelos agentes ao longo das iterações.
- Histograma `Histograma` mostra a quantidade ou popularidade de cada um dos diferentes tokens na comuniadade
- Gráfico `Fitness` mostra a distribuição de `fitness` pelos diferentes tokens.

```
turtles-own [
  token
]
globals
[
  token_values
]
to inicia
  __clear-all-and-reset-ticks
  set token_values []
  repeat num_tokens [
    ifelse aleatorio
    [set token_values lput random-normal fitness 0.2 token_values
       ]
    [set token_values lput fitness token_values]
  ]
  do-plots
  crt num_agentes
  ask turtles [
    set shape "circle"
    set token (list random num_tokens)
    set color color_token token
    setxy random-xcor random-ycor]
end

to reset
    ask turtles [
    set shape "circle"
    set token random num_tokens
    set color color_token token
```

```
    setxy random-xcor random-ycor]
end

to go
  tick
  ask turtles [
    set token shuffle token
    let t [one-of token] of one-of turtles
    while [length token > tam_memoria]
      [set token remove-item 0 token]
    ifelse (random-float 1 < novidade)
     [set token lput random num_tokens token]
     [if (random-float 1 < item t token_values)[set token lput t
        token]]
    set color color_token first token
  ]
end

to-report color_token [t]
  report int (first token / num_tokens * 140)
end

to-report entropia [lst]
  let symbols remove-duplicates lst
  let freq []
  let l (length lst)
  foreach symbols [
    let ctr 0
    let symb ?
    foreach lst [ if ? = symb [set ctr ctr + 1]]
```

```
    set freq lput (ctr / l) freq
  ]
  let rep 0.0
  foreach freq [set rep (rep + ? * log ? 2)]
  report (- rep)
end

to-report log-normal [u sgm]
  let b ln ( 1 + (sgm * sgm) / (u * u))
  let s1 sqrt(b)
  let m1 (ln(u) - (b / 2))
  report exp(random-normal m1 s1)
end

to-report tendencia [lst nivel]
  let out []
  let cnt 0
  while [length lst > 0 and cnt < nivel]
  [ foreach modes lst
    [
    set out lput ? out
    set lst remove ? lst
    set cnt (cnt + 1)
    ]]
  report one-of out
end

to do-plots
  set-current-plot "Fitness"
  let xx n-values num_tokens [?]
```

```
  clear-plot
  set-plot-x-range 0 num_tokens
  (foreach xx token_values [plotxy ?1 ?2])
end
}
```

Apêndice B

O método usual de medir a popularidade é através do valores de consumo. Os meios de comunicação social fazem-no através das métricas de audiências, as empresas a partir dos valores de vendas. Dada a facilidade com que são obtidos dados na internet usamos dados aí recolhidos. Procuramos obter uma amostra significativa de dados que pudessem garantir uma base suficiente de suporte para a validação do modelo.

Os dados foram obtidos a partir das seguintes fontes:

1. Dados recolhidos a partir do site "*Wikipedia article traffic statistics*"[1] que fornece as estatísticas diárias de consulta dos milhões de artigos existentes na Wikipedia. Para esse efeito recolhemos o numero de visitas diárias a 1963 páginas de cantores-compositores americanos, a 1781 páginas relativas a filmes americanos entre 1926 e 2006 a intervalos de uma década, e a todas as páginas dos albuns constantes no top semanal da tabela Bilboard entre 1945 e 2006 no total de 789 albuns. A popularidade de cada página foi consolidada através da média das consultas diárias durante o periodo que decorreu entre Dezembro de 2007, data a partir da qual a Wikipedia começou a disponibilizar estatísticas, e Janeiro de 2014.

2. Conjunto de dados com a popularidade de videos na rede YouTube Cha et al. [2007] [2]. Trata-se de um dataset utilizado pelos autores do artigo citado que contém a quantidade de visualizações de dois conjuntos

[1] Disponível em http://stats.grok.se/ consultado em Maio 2014.

[2] Disponível em http://an.kaist.ac.kr/traces/IMC2007.html consultado em Agosto 2014.

de 1,687,506 e 252,255 videos nas categorias de entertenimento e tecnologia e ciência recolhidos em Dezembro de 2006 e Janeiro de 2007.

3. Conjunto de dados com a evolução temporal do número de *frase-chave* e de *hashtags* emitidas num universo de texto de blogues e de *tweets* contendo 343 milhões de frases e 6 milhões de *hashtags* das quais foram retiradas 1000 como amostra Yang and Leskovec [2011][1]. O conjunto foi obtido no ano de 2009 e foi usado no estudo citado.

4. Dois conjunto de dados correspondente a uma recolha de *tweets* emitidos durante os meses que precederam as eleições presidenciais portuguesas em Janeiro de 2011 e legislativas em Junho de 2011. Os *tweets* foram emitidos por uma comunidade de 1903 utilizadores mais assiduos da rede e 44 contas de meios de comunicação social: jornais; rádios; televisões e revistas.

Um exemplo dos dados obtidos em 1 é o seguinte:

```
Album                                                  Media diaria de visitas
A_Hard_Day's_Night_(album)                             1.271364e+03
Abbey_Road                                             2.553925e+03
At_San_Quentin                                         1.681185e+02
Ballads_of_The_Green_Berets                            3.136519e+00
Beach_Boys_Concert                                     6.913670e+01
Beatles_'65                                            2.218981e+02
Beatles_VI                                             1.928128e+02
Blind_Faith_(Blind_Faith_album)                        1.498432e+02
Blood,_Sweat_&_Tears_(Blood,_Sweat_&_Tears_album)      1.842217e+01
Blue_Hawaii_(album)                                    9.284433e+01
...
```

Estes dados foram calculados a partir da recolha do número de visitas diárias por página durante todos os dias que decorreram entre 1 de Janeiro de 2007 e a primeira semana do mês de Junho de 2014. Após essa recolha os dados foram trabalhados para se obterem estas médias.

[1]Disponível em http://snap.stanford.edu/data/volumeseries.html consultado em Janeiro 2014

Um exemplo dos dados obtidos em 2 é o seguinte:

```
Video                    Visionamentos
/watch?v=lsO6D1rwrKc     8567247
/watch?v=vr3x_RRJdd4     8056321
/watch?v=ixsZy2425eY     6762465
/watch?v=RUCZJVJ_M8o     6596499
/watch?v=tFXLbXyXy6M     5371920
/watch?v=jtExxsiLgPM     5190212
/watch?v=SmLhyPjHVes     5138585
/watch?v=2KrdBUFeFtY     4112953
/watch?v=5bry10DU8bo     3793255
/watch?v=vD4OnHCRd_4     3639919
/watch?v=qg_8HIwRyrE     3521722
...
```

Estes dados foram obtidos directamente através do *dataset* fornecidos pelos autores.

Um exemplo dos dados obtidos em 3 é o seguinte:

```
hora e popularidade dos 4 perfis
1  0.01247383   0.001504988  0.001524522  0.001648017
2  0.02179988   0.003088329  0.002908063  0.002965373
3  0.03507193   0.004612104  0.003782027  0.003245257
4  0.03848731   0.005280174  0.004231336  0.003311753
5  0.03861028   0.005292791  0.004189317  0.003166232
6  0.03900322   0.005296090  0.004302741  0.002920961
7  0.03855049   0.005721875  0.004512937  0.002676595
8  0.03866987   0.006268741  0.004627809  0.002410862
9  0.03987509   0.007108101  0.004835199  0.002331603
10   0.04095550   0.008018718  0.004980255  0.002364105
11   0.04173296   0.008291741  0.005127178  0.002362445
12   0.04380656   0.008767077  0.005300880  0.002400151
...
```

Estes dados foram obtidos aplicando o algoritmo proposto no artigo ao *dataset* fornecido pelos autores com a variante da aglomeração ser efetuada em 4 *clusters* em vez de 6 no caso das palavras chave nos blogues.

Um exemplo dos dados obtidos em 4 é o seguinte:

Tweets legislativas

Data	PSD	PS	CDS	PCP	BE
2011-03-23	318	248	155	100	49
2011-03-24	703	858	196	83	62
2011-03-25	756	525	157	82	80
2011-03-26	198	216	39	9	10
2011-03-27	318	370	112	47	41
2011-03-28	203	245	65	39	18
2011-03-29	99	43	29	19	17
2011-03-30	68	17	16	36	24
2011-03-31	210	271	96	28	34
2011-04-01	257	227	92	13	51
2011-04-02	159	129	53	7	5
...					

Noticias presidenciais

Data	Cavaco	Alegre	Nobre	Lopes	Moura	Coelho
2010-10-29	1	2	0	0	0	0
2010-10-30	1	0	0	0	0	0
2010-10-31	0	3	3	2	0	0
2010-11-01	3	0	0	0	0	3
2010-11-02	1	2	1	0	0	1
2010-11-03	4	0	0	0	0	0
2010-11-04	4	0	0	0	0	0
2010-11-05	1	1	0	1	0	0
2010-11-06	0	0	0	0	0	0
2010-11-07	0	3	0	0	0	0
2010-11-08	1	2	0	1	0	0
...						

Estes dados foram obtidos a partir da contagem da menção de um conjunto de palavra chaves associadas a cada uma das entidades nos *tweets* coletados diariamente. Este processo restringiu-se a um número controlado de 1903 utilizadores e 44 contas de meios de comunicação social tendo sido controlado a existência de *outliars*.

Apêndice C

No envelope na contracapa encontra-se um CD com os seguintes conjuntos de dados e modelos:

Conjunto de dados, pasta Dados:

1. Pasta Dados Wikipedia - conforme a descrição do Apêndice B, os ficheiros contêm dados mensais retirados dos dados diários a partir dos quais são calculadas as médias diárias. Devido à dimensão dos ficheiros com as recolhas diárias optámos por incluir apenas os resumos mensais.

2. Pasta Dados Youtube - conforme a descrição do Apêndice B, os ficheiros contêm dados por cada um dos vídeos os dados de visionamento encontram-se na coluna 3. 'YoutubeEnt.txt' contém os dados da categoria de entretenimento e 'YoutubeSci.txt' contém os dados da categoria de tecnologia e ciência.

3. Pasta Dados Blogues e Twitter - conforme a descrição do Apêndice B, 'MemePhr.txt' e 'TwtHtag.txt"contêm os dados em bruto, os outros dois ficheiros contêm os centroides calculados a partir dos primeiros.

4. Pasta Dados de Eleições Twitter - conforme descrição do Apêndice B, os ficheiros 'noticias_xxx.csv' contêm os dados resumidos de notícias, os ficheiros 'tweets_xxx.csv' contêm os dados resumidos do número de *tweets*, os ficheiros 'sondagens_xxx.csv' contêm os dados das sondagens efetuadas no período. Devido ao tamanho do ficheiro com o conteúdo real dos *tweets* e todos os meta-dados optámos por incluir estas versões resumidas.

Modelos Netlogo, pasta Modelos Netlogo (os dados *batch* foram retirados de modelos adaptados para correr em modo não visual):

1. Pasta Modelo Browniano - Inclui dois ficheiros correspondentes à implementação do modelo browniano. Os sliders correspondem à parametrização do modelo. Os gráficos mostram os votos e a similaridade cosenoidal falada no texto. Nas pequenas janelas estão presentes algumas grandezas calculadas do modelo.

2. Pasta Modelo de Impacto - Inclui dois ficheiro correspondentes à implementação do modelo de impacto. A escolha 'rede' permite optar-se por percentagens de re-conexão da rede original. A escolha 'delta' permite definir o valor de δ. A escolha 'frac' permite definir a percentagens de agente que recebem notícias. O interruptor permite optar pelo calculo da similaridade em função dos *tweets* ou das notícias.

3. Pasta Modelo de Popularidade - Inclui dois ficheiros com exemplos de criação de popularidade a partir da replicação muito simples de *memes*. O funcionamento da aplicação consta do Apêndice A. A variante AlfaBetaTetha inclui uma parametrização conforme os modelos que propomos na tese. Consultar o código simples para mais detalhes.

Referências

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE, 2011. 27

Pieter Adriaans. Information. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012. 9, 11

Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. A peek into the future: predicting the evolution of popularity in user generated content. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 607–616. ACM, 2013. 27

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. 117

Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 333–352. Springer, 2013. 99

Roy M Anderson and Robert McCredie May. *Infectious diseases of humans*, volume 1. Oxford university press Oxford, 1991. 40

Werner Antweiler and Murray Z Frank. Is all that talk just noise? the information

content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004. 29

Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010. 28

Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006. 28

Roja Bandari, Sitaram Asur, and Bernardo A Huberman. The pulse of news in social media: Forecasting popularity. In *ICWSM*, 2012. 26

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999. 21

Jon Barwise. *Information flow: the logic of distributed systems*. Cambridge University Press, 1997. 15

Frank M. Bass. A New Product Growth for Model Consumer Durables. *Management Science*, 15(5):215–227, January 1969. 22

Frank M Bass. Empirical generalizations and marketing science: A personal view. *Marketing Science*, 14(3_supplement):G6–G19, 1995. 21

Slimane Ben Slimane. Bounds on the distribution of a sum of independent lognormal random variables. *Communications, IEEE Transactions on*, 49(6):975–978, 2001. 48

Rushi Bhatt, Vineet Chaoji, and Rajesh Parekh. Predicting product adoption in large-scale social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1039–1048. ACM, 2010. 28

Urs Birchler and Monica Butler. *Information Economics*. Routledge, 2007. 13

Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011a. 26

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011b. 29

Youmna Borghol, Siddharth Mitra, Sebastien Ardon, Niklas Carlsson, Derek Eager, and Anirban Mahanti. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*, 68(11):1037–1055, 2011. 24

Tim Brody, Stevan Harnad, and Leslie Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006. 27

Ronald S Burt. *Brokerage and closure: An introduction to social capital*. Oxford University Press, 2005. 122, 135

Robert D Buzzell. Are there"natural"market structures? *The Journal of Marketing*, pages 42–51, 1981. 23

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009. 96

Manuel Castells. *Rise of the Network Society*. Wiley-Blackwell, 1996. 15

Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 211–223. ACM, 2014. 27

Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM Internet Measurement Conference*, October 2007. 57, 142

Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. Analyzing the video popularity characteristics of large-scale user ge-

nerated content systems. *IEEE/ACM Transactions on Networking (TON)*, 17 (5):1357–1370, 2009. 24

Gregory J. Chaitin. Algorithmic information theory. *IBM Journal of Research and Development*, 21(4):350–359, 1977. 12

Alex Cheng, Mark Evans, and Harshdeep Singh. Inside twitter: An in-depth look inside the twitter world. *Report of Sysomos, June, Toronto, Canada*, 2009. 32

Xu Cheng, Cameron Dale, and Jiangchuan Liu. Understanding the characteristics of internet short video sharing: Youtube as a case study. *arXiv preprint arXiv:0707.3670*, 2007. 24

Xu Cheng, Cameron Dale, and Jiangchuan Liu. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238. IEEE, 2008. 24

Murphy Choy, Michelle LF Cheong, Ma Nang Laik, and Koo Ping Shung. A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *arXiv preprint arXiv:1108.5520*, 2011. 26

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009. 36, 53, 54, 64

Peter Clifford and Aidan Sudbury. A model for spatial conflict. *Biometrika*, 60 (3):581–588, 1973. 97

Thomas Couronne, Jean-Samuel Beuscart, and Cedric Chamayou. Self-organizing map and social networks: Unfolding online social popularity. *arXiv preprint arXiv:1301.6574*, 2013. 27

T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, 2006. 12

Robert T Craig. Communication theory as a field. *Communication theory*, 9(2): 119–161, 1999. 16, 17

Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008. 24

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04): 87–98, 2000. 97

Daniel Clement Dennett. *Sweet dreams: Philosophical obstacles to a science of consciousness*. MIT press, 2005. 2

Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010. 26

Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001. 28

Wenjing Duan, Bin Gu, and Andrew B Whinston. Do online reviews matter?—an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016, 2008. 28

David Easley and Jon Kleinberg. Networks, crowds, and markets. *Cambridge University*. 40

James E Engel, Roger D Blackwell, and Robert J Kegerreis. How information is used to adopt an innovation. *Journal of Advertising Research*, 9(4):3–8, 1969. 28

Jeroen Famaey, Frédéric Iterbeke, Tim Wauters, and Filip De Turck. Towards a predictive cache replacement strategy for multimedia content. *Journal of Network and Computer Applications*, 36(1):219–227, 2013. 3

Luciano Floridi. Semantic conceptions of information. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2011 edition, 2011. 11, 15, 37, 38

Dongfeng Fu, Fabio Pammolli, Sergey V Buldyrev, Massimo Riccaboni, Kaushik Matia, Kazuko Yamasaki, and H Eugene Stanley. The growth of business firms: Theoretical framework and empirical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):18801–18806, 2005. 49

Xavier Gabaix. Zipf's law for cities: an explanation. *Quarterly journal of Economics*, pages 739–767, 1999. 21

Serge Galam. Minority opinion spreading in random geometry. *The European Physical Journal B-Condensed Matter and Complex Systems*, 25(4):403–406, 2002. 97

Shuai Gao, Jun Ma, and Zhumin Chen. Popularity prediction in microblogging network. In *Web Technologies and Applications*, pages 379–390. Springer, 2014. 26

Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *ICWSM*, 2011. 33

Robert Gibrat. *Les inégalités économiques*. Recueil Sirey, 1931. 22, 45

Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *ICWSM*, pages 59–65, 2010. 29

K Goidel. Public opinion polling in a digital age: Meaning and measurement. *Political Polling in the Digital Age*, pages 11–27, 2011. 31

Gerald Joseph Goodhardt, Andrew SC Ehrenberg, and Christopher Chatfield. The dirichlet: a comprehensive model of buying behaviour. *Journal of the Royal Statistical Society. Series A (General)*, pages 621–655, 1984. 22, 23, 131

Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978. 28

A. Greenberg. Alternatives to polling. *Political Polling in the Digital Age*, pages 11–27, 2011. 32

Jakub Growiec, Fabio Pammolli, Massimo Riccaboni, and H Eugene Stanley. On the size distribution of business firms. *Economics Letters*, 98(2):207–212, 2008. 46, 49, 51, 131

Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004. 27

Daniel Gruhl, Ramanathan V. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05*, pages 78–87, 2005. 28

Manish Gupta, Jing Gao, ChengXiang Zhai, and Jiawei Han. Predicting future popularity trend of events in microblogging platforms. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012. 26

GONCA Gursun, Mark Crovella, and Ibrahim Matta. Describing and forecasting video access patterns. In *INFOCOM, 2011 Proceedings IEEE*, pages 16–20. IEEE, 2011. 93

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002. 97

Bruce M Hill. The rank-frequency form of zipf's law. *Journal of the American Statistical Association*, 69(348):1017–1026, 1974. 23

Liangjie Hong, Ovidiu Dan, and Brian D Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011. 25, 26

Yuji Ijiri and Herbert A Simon. *Skew distributions and sizes of business firms*. Amsterdam: North-Holland Pub. Co., 1977. 21, 22

Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *JASIST*, 60(11):2169–2188, 2009a. 28

Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009b. 26

Harold J Jansen and Royce Koop. Pundits, ideologues, and ranters: The british columbia election online. *Canadian Journal of Communication*, 30(4), 2005. 32

Lu Jiang, Yajie Miao, Yi Yang, Zhenzhong Lan, and Alexander G Hauptmann. Viral video style: A closer look at viral videos on youtube. In *Proceedings of International Conference on Multimedia Retrieval*, page 193. ACM, 2014. 23

Andreas Jungherr, Pascal Jürgens, and Harald Schoen. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review*, 30(2):229–234, 2012. 32

Elihu Katz and Paul Felix Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers, 1970. 28

Su-Do Kim, Sung-Hwan Kim, and Hwan-Gue Cho. Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity. In *Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on*, pages 449–454. IEEE, 2011. 27

Goidel K. Xenos M. Kirzinger, A. Too much talk, not enough action? *Political Polling in the Digital Age*, pages 11–27, 2011. 31, 99

Rajiv Kohli and Raaj Sah. *Market shares: some power law results and observations*. Irving B. Harris Graduate School of Public Policy Studies, University of Chicago, 2004. 23, 131

A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965. 12

Shoubin Kong, Ling Feng, Guozheng Sun, and Kan Luo. Predicting lifespans of popular tweets in microblog. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1129–1130. ACM, 2012. 26

Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, and Zhe Zhao. Predicting bursts and popularity of hashtags in real-time. 2014a. 26

Shoubin Kong, Qiaozhu Mei, Ling Feng, Zhe Zhao, and Fei Ye. On the real-time prediction problems of bursting hashtags in twitter. *arXiv preprint arXiv:1401.2018*, 2014b. 26

Andrey Kupavskii, Alexey Umnov, Gleb Gusev, and Pavel Serdyukov. Predicting the audience size of a tweet. In *ICWSM*, 2013. 25

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010. 127

Paul Langevin. Sur la théorie du mouvement brownien. *CR Acad. Sci. Paris*, 146(530-533), 1908. 98, 108

Bibb Latane. The psychology of social impact. *American psychologist*, 36(4):343, 1981. 98, 115

Jong Gun Lee, Sue Moon, and Kavé Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 623–630. IEEE, 2010. 27

Kristina Lerman and Aram Galstyan. Analysis of social voting patterns on digg. In *Proceedings of the first workshop on Online social networks*, pages 7–12. ACM, 2008. 30

Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97, 2010. 25

Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on World wide web*, pages 621–630. ACM, 2010. 30

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007a. 23

Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, pages 551–556. SIAM, 2007b. 29

Yu-Ru Lin, Drew Margolin, Brian Keegan, Andrea Baronchelli, and David Lazer. # bigbirds never die: Understanding social dynamics of emergent hashtag. *arXiv preprint arXiv:1303.7144*, 2013. 26

Yu-Ru Lin, Brian Keegan, Drew Margolin, and David Lazer. Rising tides or rising stars?: Dynamics of shared attention on twitter during media events. *PloS one*, 9(5):e94093, 2014. 26

JAHO Lyst, Krzysztof Kacperski, and Frank Schweitzer. Social impact models of opinion dynamics. *Annual reviews of computational physics*, 9:253–273, 2002. 116

Haixin Ma, Weining Qian, Fan Xia, Xiaofeng He, Jun Xu, and Aoying Zhou. Towards modeling popularity of microblogs. *Frontiers of Computer Science*, 7 (2):171–184, 2013a. 25

Zongyang Ma, Aixin Sun, and Gao Cong. Will this# hashtag be popular tomorrow? In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1173–1174. ACM, 2012. 26

Zongyang Ma, Aixin Sun, and Gao Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 64(7):1399–1410, 2013b. 26

Yasuko Matsubara, Yasushi Sakurai, B Aditya Prakash, Lei Li, and Christos Faloutsos. Rise and fall patterns of information diffusion: model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 6–14. ACM, 2012. 26

Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972. 103

Panagiotis Takis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (not) to predict elections. In *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*, pages 165–171. IEEE, 2011. 32, 33

Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*, 2006a. 25

Gilad Mishne and Natalie S Glance. Predicting movie sales from blogger sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 155–158, 2006b. 28

Scott Moss and Bruce Edmonds. Towards good social science. *Journal of Artificial Societies & Social Simulation*, 8(4), 2005. 96

Diana C Mutz. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press, 2006. 32

Mark Newman. *Networks: an introduction*. Oxford University Press, 2010. 117, 120

Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005. 21

Andrzej Nowak, Jacek Szamrej, and Bibb Latané. From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, 97(3):362, 1990. 97, 115

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010. 26, 29, 32, 99, 100

Kazumi Okuyama, Misako Takayasu, and Hideki Takayasu. Zipf's law in income distribution of companies. *Physica A: Statistical Mechanics and its Applications*, 269(1):125–131, 1999. 21

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010. 99

Cheol Park and Thae Min Lee. Information direction, website reputation and ewom effect: A moderating role of product type. *Journal of Business Research*, 62(1):61–67, 2009. 28

Kibeom Park and Eytan Domany. Power law distribution of dividends in horse races. *EPL (Europhysics Letters)*, 53(4):419, 2001. 21

Joseph Persky. Retrospectives: Pareto's law. *The Journal of Economic Perspectives*, pages 181–192, 1992. 21

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011. 26

Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 365–374. ACM, 2013. 27, 93

Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3, 2013. 29

Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. Characterizing and modeling the dynamics of online popularity. *Physical review letters*, 105(15):158701, 2010. 27

William J Reed and Barry D Hughes. A model explaining the size distribution of gene and protein families. *Mathematical biosciences*, 189(1):97–102, 2004. 50

J. Rissanen. Modeling by the shortest data description. *Automation*, 14:465–471, 1978. 12

Camille Roth and Jean-Philippe Cointet. Social and semantic coevolution in knowledge networks. *Social Networks*, 32(1):16–29, 2010. 118

Richard Rothenberg. The relevance of social epidemiology in hiv/aids and drug abuse research. *American journal of preventive medicine*, 32(6 Suppl):S147, 2007. 40

Alessandra Sala, Sabrina Gaito, Gian Paolo Rossi, Haitao Zheng, and Ben Y Zhao. Revisiting degree distribution models for social graph analysis. *arXiv preprint arXiv:1108.0027*, 2011. 117

José Carlos Santos and Sérgio Matos. Predicting flu incidence from portuguese tweets. In *IWBBIO*, pages 11–18, 2013. 27

Thomas C Schelling. *Micromotives and macrobehavior*. WW Norton & Company, 2006. 96

Frank Schweitzer. *Brownian agents and active particles: collective dynamics in the natural and social sciences*. Springer, 2007. 98, 108

Frank Schweitzer and David Garcia. An agent-based model of collective emotions in online communities. *The European Physical Journal B*, 77(4):533–545, 2010. 97, 108, 109

David Shamma, Lyndon Kennedy, and Elizabeth Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events. *CSCW Horizons*, 2010. 26

David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, pages 3–10. ACM, 2009. 26

C. E. Shannon and W. Weaver. *The mathematical theory of communication*. The University of Illinois Press, Urbana, IL, 1949. 11

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948. 11

Ramesh Sharda and Dursun Delen. Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2):243–254, 2006. 28

Clay Shirky. *Here comes everybody: The power of organizing without organizations*. Penguin, 2008. 31

Clay Shirky. Political power of social media-technology, the public sphere sphere, and political change, the. *Foreign Aff.*, 90:28, 2011. 31

Herbert A Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955. 21

Herbert A Simon and Charles P Bonini. The size distribution of business firms. *The American Economic Review*, pages 607–617, 1958. 21, 22

Jeffrey S Simonoff and Ilana R Sparrow. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*, 13(3):15–24, 2000. 28

John J Skowronski and Donal E Carlston. Negativity and extremity biases in impression formation: A review of explanations. *Psychological bulletin*, 105(1): 131, 1989. 28

Pawel Sobkowicz. Modelling opinion formation with physics tools: Call for closer link with reality. *Journal of Artificial Societies & Social Simulation*, 12(1), 2009. 96

R. J. Solomonoff. A preliminary report on a general theory of inductive inference. 1960. 12

Michael HR Stanley, Sergey V Buldyrev, Shlomo Havlin, Rosario N Mantegna, Michael A Salinger, and H Eugene Stanley. Zipf plots and the size distribution of firms. *Economics letters*, 49(4):453–457, 1995. 21

Carl W Stern and Michael S Deimler. *The Boston consulting group on strategy: Classic concepts and new perspectives*. John Wiley & Sons, 2006. 23

John Sutton. Gibrat's legacy. *Journal of economic literature*, pages 40–59, 1997. 22, 46, 131

Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010. 25, 26, 30

Katarzyna Sznajd-Weron and Jozef Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000. 97

Alexandru Tatar, Jérémie Leguay, Panayotis Antoniadis, Arnaud Limbourg, Marcelo Dias de Amorim, and Serge Fdida. Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 67. ACM, 2011. 27

Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. Ranking news articles based on popularity prediction. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 106–110. IEEE Computer Society, 2012. 25

Alexandru Tatar, Marcelo Dias de Amorim, Serge Fdida, and Panayotis Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1):1–20, 2014. 24, 93

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011. 26

Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. News comments: Exploring, modeling, and online prediction. In *Advances in Information Retrieval*, pages 191–203. Springer, 2010. 25

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010. 30, 32, 99, 100

Thomas W Valente and Patchareeya Pumpuang. Identifying opinion leaders to promote behavior change. *Health Education & Behavior*, 2007. 40

P Van Mieghem, N Blenn, and C Doerr. Lognormal distribution in the digg online social network. *The European Physical Journal B-Condensed Matter and Complex Systems*, 83(2):251–261, 2011. 25

Jean Véronis. Citations dans la presse et résultats du premier tour de la présidentielle 2007. *Retrieved December*, 15:2009, 2007. 100

Chunyan Wang, Mao Ye, and Bernardo A Huberman. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 244–252. ACM, 2012. 25

Duncan J Watts. The"new"science of networks. *Annual review of sociology*, pages 243–270, 2004. 120

Warren Weaver. The mathematics of communication. *Communication theory*, pages 27–38, 1949. 13

Christine Williams and Girish Gulati. What is a social network worth? facebook and vote share in the 2008 presidential primaries. In *Annual Meeting of the American Political Science Association*, volume 54, 2008. 26, 32

Christine B Williams and Girish J Gulati. Social networks in political campaigns: Facebook and the 2006 midterm elections. In *annual meeting of the American Political Science Association*, 2007. 26

Mauro Wolf. *Teorie delle comunicazioni di massa*. Bompiani Milano, 1985. 17, 18

Mauro Wolf and Maria Jorge Vilar de Figueiredo. *Teorias da comunicação*. Presença, 1987. 103

Kazuko Yamasaki, Kaushik Matia, Sergey V Buldyrev, Dongfeng Fu, Fabio Pammolli, Massimo Riccaboni, and H Eugene Stanley. Preferential attachment and growth dynamics in complex systems. *Physical Review E*, 74(3):035103, 2006. 49

Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186. ACM, 2011. 27, 77, 78, 143

Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. *ICWSM*, 10:355–358, 2010. 25

Sheng Yu and Subhash Kak. A survey of prediction using social media. *CoRR*, abs/1203.1647, 2012. 27

Tauhid Zaman, Emily B Fox, and Eric T Bradlow. A bayesian approach for predicting the popularity of tweets. *arXiv preprint arXiv:1304.6777*, 2013. 25

Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, volume 104, pages 17599–601. Citeseer, 2010. 30

Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 301–304. IEEE Computer Society, 2009. 28

Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011. 29

George Kingsley Zipf. Human behavior and the principle of least effort. 1949. 21

George Kingsley Zipf. A note on brand-names and related economic phenomena. *Econometrica, Journal of the Econometric Society*, pages 260–263, 1950. 21