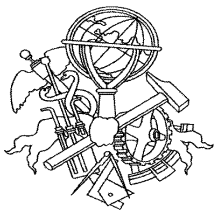
# A LEI DE ZIPF OU REGRA 80/20 DA INTERNET

Hélvio Sandro Vieira Bosso



Mestrado em Engenharia Eletrotécnica e de Automação
Área de Especialização de Automação e Sistemas
Departamento de Engenharia Electrotécnica
Instituto Superior de Engenharia do Porto
2012

Este relatório satisfaz, parcialmente, os requisitos que constam da Ficha de Disciplina de Tese/Dissertação, do 2º ano, do Mestrado em Engenharia Eletrotécnica e de Computadores

Candidato: Hélvio Sandro Vieira Bosso, Nº 1091444, 1091444@isep.ipp.pt
Orientação científica: Carla Pinto, cap@isep.ipp.pt
Coorientação Cientifica: J. A. Tenreiro Machado, jtm@isep.ipp.pt, António Mendes Lopes, aml@fe.up.pt

Mestrado em Engenharia Eletrotécnica e de Computadores
Área de Especialização de Automação e Sistemas
Departamento de Engenharia Eletrotécnica
Instituto Superior de Engenharia do Porto
23 de Julho de 2012

# *Agradecimentos*

Esta jornada começou como um desafio, mas com o passar do tempo tornou-se um objetivo e uma meta a cumprir.

Agradeço a Deus por conduzir sempre os meus caminhos, nos momentos mais difíceis mostrar-me a solução e a decisão mais sensata a ser tomada.

Agradeço à minha mãe Teresa Bessa pela educação e a experiência passada, ao meu Pai Luís Bosso pelos valores e o seu bom coração, as irmãs Nádia, Sofia e a Délcia, namorada Ludmila Vaz e os amigos cujo empenho, carinho e solidariedade permitiram-me concluir com sucesso esta etapa que encerra mais um ciclo académico.

No meio académico, quero expressar o meu agradecimento à professora Carla Pinto, pelo apoio incondicional demonstrado, desde o início até ao final deste trabalho, e aos professores António Lopes e Tenreiro Machado agradeço também toda a disponibilidade demonstrada.

# *Resumo*

As Leis de Potência, LP, (*Power Laws*, em inglês), Leis de Pareto ou Leis de Zipf são distribuições estatísticas, com inúmeras aplicações práticas, em sistemas naturais e artificiais. Alguns exemplos são a variação dos rendimentos pessoais ou de empresas, a ocorrência de palavras em textos, as repetições de sons ou conjuntos de sons em composições musicais, o número de vítimas em guerras ou outros cataclismos, a magnitude de tremores de terra, o número de vendas de livros ou CD's na internet, o número de sítios mais acedidos na Internet, entre muitos outros. Vilfredo Pareto (1897-1906) afirma, no manual de economia política "*Cours d'Economie Politique*", que grande parte da economia mundial segue uma determinada distribuição, em que 20% da população reúne 80% da riqueza total do país, estando, assim uma pequena fração da sociedade a controlar a maior fatia do dinheiro. Isto resume o comportamento de uma variável que segue uma distribuição de Pareto (ou Lei de Potência).

Neste trabalho pretende-se estudar em pormenor a aplicação das leis de potência a fenómenos da internet, como sendo o número de sítios mais visitados, o número de links existentes em determinado sítio, a distribuição de nós numa rede da internet, o número livros vendidos e as vendas em leilões online. Os resultados obtidos permitem-nos concluir que todos os dados estudados são bem aproximados, numa escala logarítmica, por uma reta com declive negativo, seguindo, assim, uma distribuição de Pareto.

O desenvolvimento e crescimento da *Web*, tem proporcionado um aumento do número dos utilizadores, conteúdos e dos sítios. Grande parte dos exemplos presentes neste trabalho serão alvo de novos estudos e de novas conclusões. O fato da internet ter um papel preponderante nas sociedades modernas, faz com que esteja em constante evolução e cada vez mais seja possível apresentar fenómenos na internet associados Lei de Potência.

***Palavras-Chave***

Leis de Potência, Lei de Zipf, Lei de Pareto, e-mail, internet, links, páginas, sítio, utilizadores.

# *Abstract*

Power laws are statistical distributions. They have been applied to the modeling of various phenomena of everyday life. Some examples are the variation in personal incomes or businesses, the occurrence of words in texts, the repetition of sounds or combinations of sounds in musical compositions, the number of casualties in wars or other cataclysms, the magnitude of earthquakes, the number of sales of books or CD's on the internet, the number of sites on the Internet, among many others. Vilfredo Pareto (1897-1906) states, in his manual of political economy, "Cours d'Economie Politique", that a large part of the world economy follows a particular distribution, in which 20% of the population gathers 80% of the total wealth, and so, a small fraction of the society controls the largest share of the money. This summarizes the behavior of a variable that follows a Pareto distribution (or, Power Law).

This work intends to study in detail the application of power laws (PL) to the internet phenomena, for example, the number of top sites, the number of links on a particular site, the distribution of internet nodes, the number of books sold online, online auctions. The results obtained allow us to conclude that all data studied are well approximate, on a logarithmic scale, by a straight line with a negative slope, thus following a Pareto distribution.

The development and growth of the Web, has provided an increasing number of users, content and sites. Most of the examples in this work will be the target of new studies and new findings. The fact that the internet has a preponderant role in modern societies, forces its constant evolution and increasing, and it is possible that new internet phenomena associated with PL may appear.

***Keywords***

Power Law, Zipf Law, Pareto Law, e-mail, internet, links, webpages, sites, users.

# *Índice*

# *Índice de Figuras*

# *Índice de Tabelas*

# *Acrónimos*

FDP – Função Densidade da Probabilidade

LP – Lei de Potência

P2P – Peer-to-Peer

ISP – Internet Service Providers

WWW – World Wide Web

# 1. INTRODUÇÃO

As Leis de Potência, LP, (*Power Laws*, em inglês), Leis de Pareto ou Leis de Zipf são distribuições estatísticas, com inúmeras aplicações práticas, em sistemas naturais e artificiais. Alguns exemplos são a variação dos rendimentos pessoais ou de empresas, a ocorrência de palavras em textos, as repetições de sons ou conjuntos de sons em composições musicais, o número de vítimas em guerras ou outros cataclismos, a magnitude de tremores de terra, o número de vendas de livros ou CD's na internet, o número de sítios mais acedidos na Internet, entre muitos outros.

Vilfredo Pareto (1897-1906), publicou, em 1896, o manual de economia política "*Cours d'Economie Politique*" [8]. No mesmo defende que grande parte da economia mundial segue uma determinada distribuição, denominada, mais tarde, de "Lei de Pareto". Esta lei é também conhecida pela "Lei 80/20", que significa que 20% da população reúne 80% da riqueza total do país, estando, assim uma pequena fração da sociedade a controlar a maior fatia do dinheiro.

Ao longo das últimas décadas, a Lei de Pareto tem sido alvo de muitos estudos por parte de vários investigadores e cientistas de todas as áreas do saber. Estoup [21] observou que a frequência das palavras que são utilizadas num texto parece seguir uma LP. Esta sua observação foi analisada mais profundamente e confirmada por Zipf [48], [47]. Price [42] estudou o número de citações de artigos científicos e verificou que o número de citações recebidas aparenta ter uma distribuição semelhante a uma LP. Alice Hackett, editora da Publisher's Weekly estudou a distribuição acumulada do número total de exemplares vendidos nos Estados Unidos referente aos 633 livros considerados *best-sellers* (isto é, que

venderam 2 milhões ou mais de cópias, entre 1895 e 1965) [27]. Os dados foram compilados meticulosamente durante um período de várias décadas e o livro mais vendido foi o de Benjamin Spock, “The Common Sense Book of Baby and Child Care”. Aiello [10] observou a existência de uma distribuição cumulativa no número de chamadas recebidas num dia, nos Estados Unidos da América, por 51 milhões de utilizadores da companhia de telefone AT & T. O maior número de chamadas recebidas por um cliente naquele dia foi 375 746, ou seja cerca de 260 chamadas por minuto (tratava-se de um “call center”). Distribuições similares são observadas para o número de chamadas realizadas pelos utentes e também para o número de mensagens de e-mail que as pessoas enviam e recebem [20], [30]. A distribuição cumulativa da magnitude, medida na escala de Ritcher, dos terramotos ocorridos na Califórnia, entre de 1910 e Maio de 1992, é uma lei de potência. Esta LP traduz uma relação entre a amplitude e a frequência de ocorrência de terramotos, que é representada graficamente, em escala logarítmica, por uma reta, com declive negativo. A magnitude de Richter é definida como sendo o logaritmo, na base 10, da amplitude máxima do movimento detetada num terramoto. No gráfico da LP, aparece o logaritmo da amplitude no eixo horizontal. Neukum e Ivanov [39] observaram que o diâmetro das crateras da Lua segue uma distribuição de LP. Os autores mediram o número de crateras de um determinado tamanho em toda a superfície da Lua.

As cidades são sistemas complexos que diferem entre si no tamanho, forma e escala. O seu estudo tem sido alvo de investigação em várias áreas da ciência, desde a geografia à economia [11]. Um dos fatores que atraiu mais atenção foi a distribuição do tamanho (número de indivíduos da população) das cidades. Existem duas distribuições que regulam o tamanho das cidades: a distribuição de Pareto (que inclui a Lei de Zipf, como caso particular) e a distribuição log-normal. Krugman [32] estudou (1991) 135 áreas metropolitanas dos EUA e calculou que, numa escala logarítmica, o tamanho das cidades e a sua frequência é bem aproximado por uma reta com declive -1. Gabaix [24], [25] observou que utilizando os mesmos dados derivou uma justificação estatística para a Lei de Zipf. Assim sendo se as cidades seguirem processos de crescimento similares, i.e., se evoluírem com a mesma média e com a mesma variância, para um determinado intervalo (normalizado) de tamanhos, no estado estacionário, a distribuição do tamanho das cidades seguiria uma Lei de Zipf. Processos de crescimentos semelhantes são frequentemente designados de Lei de Gibrat [26] ou Lei de Gibrat de Efeitos Proporcionais. Desta forma a

Lei Zipf é considerada um estado estacionário da Lei de Gibrat, isto é a Lei de Zipf é o limite de um processo estocástico.

Cederman [15] estudou um modelo para o número de vítimas de uma guerra com base no trabalho realizado por Richardson [43], [44]. Este modelo assume que o conflito se expande e difunde potencialmente, durante longos períodos de tempo, devido à sua execução quase paralela. Este foi um desenvolvimento relativamente aos modelos anteriores, pois permitiu modelar dinamicamente as entidades territoriais com fronteiras flutuantes. Este aspeto foi bastante importante dado que as grandes guerras mundiais coincidiram com acentuadas reformulações do mapa geopolítico.

Clauset *et al* [18] estudaram a frequência e a gravidade dos ataques terroristas desde 1968, e concluíram que o gráfico em escalas logarítmicas da frequência em função da gravidade dos ataques seguia uma distribuição LP. A gravidade dos ataques foi medida por três fatores, número de mortes, lesões e a soma dos dois anteriores. Fatores como o desenvolvimento económico, tipo de arma utilizada durante a realização do ataque, ou o tempo de duração não tiveram qualquer influência nos resultados obtidos. Foi também estimada a periodicidade do terrorismo mundial, com um valor aproximado de 13 anos. Como conclusão, os autores propuseram um modelo matemático para a frequência dos ataques terroristas e demonstraram que modelos anteriores não conseguiam prever o comportamento da distribuição LP.

## 1.1. CONTEXTUALIZAÇÃO

Nas últimas décadas assistimos ao nascimento e crescimento exponencial dos conteúdos e utilizadores na WWW (*World Wide Web*). A utilização da internet alterou de forma drástica o nosso quotidiano, ao nível do processo de comunicação, recolha de informação, condução dos negócios e da realização de compras. Em 1996 eram 61 milhões de utilizadores e em 2000 estes valores atingiram os 400 milhões [6].

Adamic e Huberman [4] estudam o crescimento do número de sítios na internet. A popularidade dos sítios segue uma determinada tendência que pode ser explicada matematicamente, por conseguinte, pode-se também prever o seu comportamento no futuro. A forma matemática é uma distribuição estatística, denominada de “Lei de Potência” (LP). Um caso particular desta LP é a Lei de Zipf.

Neste trabalho estuda-se a aplicação das LPs a fenómenos da internet, como o número de utilizadores de sítios, o número de links de sítios, entrega de conteúdos na *Web*, acessos ao *backbone*, propagação de vírus, venda de livros, utilização da rede de e-mail, vendas em leilões online, entre outros.

## 1.2. OBJETIVOS

O objetivo principal deste trabalho consiste no estudo das Leis de Potência (LP) e na sua aplicação a fenómenos relacionados com a internet.

Os pontos fundamentais tratados nesta tese são:

- Resumo sobre o aparecimento, desenvolvimento e o crescimento da internet;
- Descrição sucinta das LPs e, em particular, da Lei de Zipf e suas aplicações mais comuns;
- Aplicação da LP a fenómenos relacionados com a internet.

## 1.3. CALENDARIZAÇÃO

A calendarização das várias fases de desenvolvimento do trabalho é apresentada na Tabela 1. As tarefas desenvolvidas consistiram no estudo de bibliografia, que teve como base artigos científicos na área; na obtenção e análise de informação existente em bases de dados públicas e de confiança, sobre a internet; e, por fim, na elaboração do relatório final. O estudo de bibliografia foi a tarefa que se prolongou mais no tempo, dada a falta de conhecimento do mestrando sobre este assunto.

**Tabela 1 Calendarização do Projeto**

| | | Out. | | | | Nov. | | | | Dez. | | | | Jan. | | | | Fev. | | | | Mar. | | | | Abr. | | | | Mai. | | | | Jun. | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ETAPAS | SEMANAS | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Recolha da Bibliografia | 15 S | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Estudo da Bibliografia | 11 S | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Implementação | 8 S | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Analise dos dados | 12 S | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Elaboração do Relatório Final | 24 S | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

## 1.4. ORGANIZAÇÃO DO RELATÓRIO

No Capítulo 1, é feita uma introdução às leis de potência. No capítulo seguinte, é apresentada uma descrição mais detalhada da LP e as suas principais aplicações relacionadas com a internet. No terceiro capítulo, é exposto o resultado das aplicações através das LPs, gráficos das distribuições cumulativas dos fenómenos estudados. Por último, no quarto capítulo são reunidas as principais conclusões e perspetivam-se futuros desenvolvimentos.

# 2. LEIS DE POTENCIA E SUA APLICAÇÃO À INTERNET

Nas últimas décadas assistimos ao nascimento e crescimento exponencial, ao nível dos conteúdos e utilizadores, da *World Wide Web (WWW)*. A utilização da internet alterou de forma drástica o nosso quotidiano ao nível do processo de comunicação, recolha de informação, condução dos negócios e da realização de compras. Em 1996 existiam 61 milhões de pessoas que utilizavam a internet. No final de 1998 o número de utilizadores estava acima dos 147 milhões e em 2000 estes valores duplicaram para 400 milhões [6].

## 2.1. LEI DE POTENCIA

A Lei de Potência aplicada a um dado sistema permite estabelecer os pressupostos seguintes:

- apresenta características comuns de um sistema dinâmico não linear, o sistema encontra-se num estado caótico e próximo da organização automática;
- existe a possibilidade de estabelecer uma união entre diferentes tamanhos e comprimentos com base na auto-similaridade de grande e pequena dimensão;

- o facto de a LP ser comum para todas as dimensões demonstra a auto-consistência interna na forma geométrica complexa e a sua unidade em todos os limites.

A distribuição cumulativa de uma variável *X* que segue uma LP é definida como:

$$P_r(X \geq x) = (\frac{k}{x})^{\alpha} \qquad (1)$$

onde *x* é um valor no intervalo definido para *X*, k ≥ 0 é um parâmetro denominado de parâmetro de localização e $\propto \geq 0$ é o parâmetro de Pareto. Num gráfico, numa escala logarítmica nos dois eixos, $\propto$ é o declive da reta que melhor aproxima o conjunto de dados.

A função densidade de probabilidade (FDP) da variável *X* pode ser obtida a partir da derivada da equação (1):

$$f(x) = \propto k^{\propto} x^{-\propto -1} \qquad (2)$$

A média de *f*(*x*) é infinita para α ≤ 1 e a variância é também infinita para α ≤ 2.

Quando *X* é uma variável aleatória discreta, a função de probabilidade é inversamente polinomial, dada por:

$$P_r(X = x) = C\,\frac{1}{x^{\propto+1}} \qquad (3)$$

Aplicando logaritmos à equação (3) vem:

$$\ln\big(P_r(X = x)\big) = (-\propto -1)\ln x + \ln \mathrm{C} \qquad (4)$$

A equação (4) é utilizada para visualizar a FDP e para derivar α utilizando uma aproximação pelo método dos mínimos quadrados, reduzindo o quadrado dos resíduos.

## 2.2. APLICAÇÃO À INTERNET

A *World Wide Web* e o e-mail impulsionaram diversos estudos, por parte de físicos e cientistas ligados às áreas da computação e da informática. Inicialmente, os cientistas depararam-se com uma grande variedade ao nível de conteúdos, tamanho e características distintas nos fenómenos relacionados com a internet. Todavia, rapidamente se aperceberam da existência de um padrão generalizado nas medições realizadas. Observaram a existência de muitos elementos pequenos contidos na *Web* e poucos de grande dimensão. Por exemplo, alguns sítios são constituídos por milhões de páginas, mas é possível encontrar milhões de sítios que contêm um número reduzido de páginas. Um outro exemplo é o facto de alguns sítios conterem milhões de *links*, mas muitos sítios possuírem apenas um ou dois *links*. Foi também comprovado que a distribuição dos utilizadores que acedem a determinado sítio segue uma LP universal, o que leva a concluir que um número pequeno de sítios domina o tráfego na internet de um grande segmento da população. A distribuição desproporcional, do volume de utilizadores entre os diversos sítios, revela que o vencedor fica com todo o mercado e está no topo [23].

A concentração de visitantes em alguns sítios não está unicamente associada ao facto dos utilizadores considerarem esse sítio mais interessante do que outros [6]. Este facto verificou-se com a realização de uma análise a duas categorias de sítios: a primeira com informação direcionada para os adultos e a segunda dentro do domínio "edu" (educação). No primeiro caso, a informação disponibilizada possui uma seleção ao nível dos conteúdos (imagens e opção de vídeo ou *chat*). No segundo caso é possível encontrar os seguintes dados: informações académicas e de pesquisa, informações pessoais dos alunos, e funcionários, informações do pessoal do corpo docente. Este tipo de informação é do interesse do público em geral. Os resultados mostraram que a distribuição de visitas aos sítios foi desigual. O sítio mais procurado direcionado a adultos, contabilizou 1.4% do volume de visitas, enquanto 10% dos sítios englobaram 60% do volume. Observou-se o mesmo comportamento para as visitas dos sítios de domínio edu. O sítio mais visitado, umich.edu (universidade de Michigan), obteve 2,81% do volume de visitas, enquanto os 5% dos sítios mais visitados contabilizaram mais de 60% do tráfego dos visitantes. Os resultados obtidos são de interesse para os economistas e para os fornecedores. Os primeiros estudam a eficiência dos mercados no comércio eletrónico, e os segundos observam o número de clientes que determinado negócio poderá atrair. Um sítio criado

recentemente tem uma elevada probabilidade de atrair um pequeno número de visitantes e uma probabilidade muito menor de atrair um número muito grande de visitantes.

As LPs também podem ser observadas no número de *links* por sítio. Este fenómeno revela-se de duas formas. De um lado do *link* está o sítio origem e do outro lado está o sítio de receção. A popularidade de um sítio pode ser medida pelo número de ligações que tem. Ou seja, quanto mais utilizadores acederem a um determinado sítio, mais conhecido o mesmo se torna e mais *links* recebe. O crescimento do número de acessos externos de um sítio é semelhante ao crescimento de um sítio em termos do número de páginas que ele contém. Alguns sítios têm a possibilidade de adicionar rapidamente páginas ou conteúdos como diretórios ou o índice das páginas, enquanto outros realizam este processo mais lentamente e fornecem um conteúdo preferencial que aponta para outros recursos. A existência de uma diferença acentuada no número de acessos externos de um sítio conduz a um fenómeno denominado de "*small world*". Enquanto alguns sítios têm predominantemente poucos *links* para páginas de conteúdo semelhante, um elevado número de sítios são ligados por diretórios e índices, que têm milhares de *links*. Desta forma, é necessário navegar, em média, por 4 sítios para se movimentar de uma página para outra [1]. Ao nível de uma página da internet, são necessários, no máximo, 19 *links* para passar dessa página para qualquer outra [7]. Huberman e Adamic [29], [2] descrevem que o número de *links* que um nó pode receber num intervalo de tempo é proporcional a uma fração aleatória do número de ligações recebidas pelo nó.

A rede de e-mail é definida como um gráfico direcionado, onde cada e-mail representa um link que estabelece uma ligação entre o remetente e o destinatário, estas ligações são geridas por diferentes processos. A distribuição dos nós é aleatória e isto implica que todos os nós têm a mesma probabilidade de receber um e-mail, esta particularidade pode ser descrita por uma LP.

A Lei de Zipf é um caso especial de uma LP, com parâmetro $\alpha \approx 1$. A Lei de Zipf começou por ser aplicada ao tamanho das cidades. Observa-se um elevado número de cidades pequenas em comparação com um número menor de grandes [44]. Outras aplicações desta lei são, por exemplo, o estudo da frequência de palavras num texto [38].

Diversos estudos têm sido realizados para explicar a estrutura de entrega dos conteúdos na Internet, segundo a Lei de Zipf. As empresas gestoras de serviços na Internet (ISP's)

procuram desenvolver soluções para suportar e responder ao rápido crescimento do tráfego na Internet, mantendo a qualidade dos seus serviços e dando resposta às solicitações de ficheiros, realizadas pelos utilizadores. A constante solicitação de ficheiros às ISP despoletou a necessidade de serem colocadas cópias dos ficheiros em memória cache junto à localização dos utilizadores. Esta implementação é bastante útil sempre que são solicitados dados a um servidor que esteja fora do espaço geográfico do utilizador. Neste processo é utilizado um servidor *proxy* com a função de enviar os ficheiros solicitados por determinado utilizador e armazenar os mesmos localmente na cache. As solicitações de outros utilizadores serão enviadas diretamente para essa cache, sem haver uma ligação ao servidor remoto. As principais vantagens da implementação deste processo são visíveis na redução do tempo de resposta, maior largura da banda, redução do tráfego na rede e eliminação das transferências excessivas em cada *link*. Para as ISP's representa uma diminuição dos custos [5]. Cunha e Breslau [19], [12] concluíram que a popularidade dos ficheiros solicitados na Internet segue uma distribuição de Zipf e que, na memória cache, só devem ser armazenados os ficheiros mais solicitados. Desta forma é possível dar resposta a todos os pedidos dos utilizadores.

A Internet é constituída por redes em vários níveis. A WWW é uma rede de páginas interligadas. O *backbone*[1] é uma rede física que transmite dados, incluindo páginas da Web, sob a forma de pacotes, de um local para outro [19].

Adamic e Jeong [1], [31] realizaram medições na WWW e Faloutsos e Albert [22], [9] no *backbone* e ficou demonstrado que não seguem o modelo clássico de Erdos-Renyi para gráficos aleatórios. O modelo tradicional de Erdos-Renyi apresenta uma distribuição dos nós baseados na distribuição de Poisson. A maioria dos nós apresenta um número característico de links. A WWW e o *backbone* seguem uma distribuição aproximada à lei de Zipf dada por: $P \approx k^{-\tau}$, onde k representa o grau do nó e $\tau$ é o expoente. A lei de Zipf do *backbone* implica que alguns nós da rede têm um grande número de ligações, enquanto a maioria tem apenas uma ou duas ligações. Este fato traduz-se num problema se os nós que tem muitas ligações forem afetados. Se forem retirados nós com grau elevado, a rede pode ficar comprometida. Para evitar este problema, foram desenvolvidas recentemente

[1] Backbone – é definido como sendo o esquema de ligações centrais de um sistema mais amplo, tipicamente de elevado desempenho.

comunicações peer-to-peer (P2P). Nestas redes P2P, milhões de utilizadores trocam ficheiros diretamente entre si. A distribuição do número de computadores a que um computador está ligado é também a lei de Zipf. Estas redes não possuem um servidor central, pelo que cada utilizador envia o seu pedido a todos os seus vizinhos e assim sucessivamente. O inconveniente é que isto pode facilmente congestionar a rede. Adamic [3] descobriu que enviar os pedidos para os nós de maior grau pode ajudar a descongestionar a rede.

Uma das principais atividades dos utilizadores da *Web* é seguir *links* até atingir a página pretendida. Este processo pode ser longo e o utilizador acabar por se perder no processo de navegação [34], [35]. Kemeny e Snell [33] propõem um modelo estocástico com vários caminhos ou sequências de *links*. Em cada *link* o utilizador pode optar por seguir um *link* a partir da página *Web* em que se encontra ou parar a navegação. Cada escolha de um *link* tem associada uma probabilidade que pode ser interpretada como a relevância de seguir aquele *link*. Huberman e Lukose [28], [36] apresentaram o seu modelo de navegação para os utilizadores com base no movimento de Browniano[2]. Em cada passo da navegação, o "valor" que um utilizador obtém, pela escolha de um determinado *link* e navegação para outra página, é modelado por uma variável aleatória normal, independente e identicamente distribuída. A navegação continua se o valor esperado por escolher um *link* é maior do que zero. A navegação termina, caso contrário. Huberman *et al* [28] demonstram que a probabilidade de se percorrer um caminho de comprimento t é aproximadamente $t^{-\frac{3}{2}}$, que é uma lei de Pareto.

Levene *et al* [34], [35] demonstram a existência da lei de Zipf no comportamento dos surfistas da *Web*. Os utilizadores apresentam uma maior probabilidade de optar por caminhos mais rápidos. Estes caminhos são facilmente encontrados, porque são em muito menor número do que caminhos maiores. Em termos do modelo Browniano pode dizer-se que o "valor" obtido pelo utilizador em caminhos curtos é maior do que em caminhos longos.

[2] Browniano - movimento irregular de partículas num fluido originado pelas colisões entre as moléculas do fluido e as partículas.

A propagação de vírus ou *worms* a partir de e-mails enviados e recebidos é um problema atual e tem infetado milhares de computadores em todo mundo. Através de um programa anexado a um e-mail é possível infetar um computador, e em seguida todos os contatos da lista do destinatário irão receber um e-mail infetado. Deste modo a propagação é realizada em várias direções. Este processo causa elevados prejuízos às redes de computadores, destrói os dados contidos no disco rígido e pode provocar a sobrecarga dos servidores de e-mail. Numa rede organizada a informação expande-se facilmente, este fator facilita a elevada taxa de infeção e propagação do vírus, mas o mesmo não acontece se a rede for aleatória ou desorganizada, onde em alguns casos o vírus desaparece [41]. Por exemplo, em Maio de 2000, o vírus "*I love you*" infectou mais de 500000 computadores em todo mundo e obstruiu 21% das redes de computadores [17]. Uma rede de e-mails, onde cada nó seja um endereço de e-mail e cada ramo represente as mensagens trocadas entre esses endereços, segue uma LP. Este facto facilita a persistência e a propagação dos vírus. Desta forma, há medidas que devem ser tomadas, como, por exemplo, a monitorização mais frequente de nós que tenham grau elevado (isto é, muitos ramos). Com esta medida pode-se prevenir os danos causados por e-mail com vírus.

## 2.3. OUTROS CASOS DE APLICAÇÕES DE PL

Nos últimos anos, diversos empresários do ramo de comércio eletrónico, têm tipo bastantes lucros nas suas vendas, devido ao fato de realizarem vendas online de produtos bastantes procurados no mercado tais como, livros, filmes e CDs de música.

Chevalier *et al* [16] estudaram e avaliaram os dados referentes às vendas online de 18.000 livros por parte de livrarias (*Amazon* e *Barnes* & *Noble*). O número de vendas dos dois fornecedores, foi modelado por uma distribuição de Pareto com coeficiente $\alpha \approx 1.2$.

Em 2003, Brynjolfsson *et al* [13] ajustaram uma distribuição de Pareto às vendas de títulos obscuros da *Amazon.com*. O coeficiente de Pareto obtido foi $\alpha \approx 1.5$. Baseados neste valor e assumindo que os 100000 títulos mais populares são vendidos no comércio mais tradicional, estes autores concluíram que 40% das vendas da *Amazon.com* eram títulos que não podiam ser comprados nestas lojas.

Newman [40], em 2005, usou dados de 663 mais vendidos na América, desde 1985 até 1965, para mostrar que o número de cópias vendidas seguia uma distribuição de Pareto. Ele estimou a proporção de vendas usando técnicas de regressão. Newman calculou ainda

as distribuições cumulativas de 12 quantidades diferentes de sistemas físicos, biológicos, tecnológicos e sociais, tais como a frequência de palavras, chamadas telefónicas, magnitude de terramotos, citações de artigos científicos, etc. O gráfico em dupla escala logarítmica destes dados também revelou uma lei de Pareto. Os coeficientes de Pareto nesses exemplos foram calculados, estando no intervalo [0.94, 2.5].

Os leilões online têm vindo a crescer acentuadamente, devido ao fato de proporcionarem operações rápidas, mais baratas e cómodas, comparativamente a outras formas tradicionais. Associado a este fato, a internet permitiu aumentar o número de participantes dos leilões e superar barreiras associadas às restrições geográficas e o tempo [45]. O *eBay* é líder de mercado nesta área e é considerado o maior sítio de leilões do mundo, registando mais de 40 milhões de consumidores. Num espaço de cinco anos, as suas receitas cresceram mais de 100.000% [37], [14]. Yang *et al* [46] estudam dados de uma licitação feita por centenas de milhares de agentes e concluem que existe um comportamento que provoca algo inesperado sobre todos os acontecimentos relacionados com o leilão (desde as licitações feitas pelos agentes que participam no leilão até ao preço final a que o produto será vendido). O número total de ofertas colocadas numa única categoria, por determinado agente, segue uma LP. O comportamento da LP permite prever que o agente que realiza licitações com maior frequência, tem maior probabilidade de, no instante seguinte, voltar a licitar, para além de exercer grande influência perante os outros agentes e no preço final do produto. O número de itens distintos que são visitados por determinado agente também segue uma distribuição de Pareto. Esta distribuição implica que poucos agentes licitam mais frequentemente e em mais itens distintos do que outros. Os leilões online são conduzidos por processos auto-organizados que reúnem todos os participantes do evento.

# 3. APLICAÇÃO DAS LEIS DE POTENCIA A DADOS REAIS DA INTERNET

Neste capítulo aplicam-se as Leis de Potência a dados reais de fenómenos relacionados com a internet. Estuda-se a distribuição cumulativa das variáveis associadas a cada fenómeno.

A aplicação de uma LP a um conjunto de dados da internet segue, habitualmente, o procedimento abaixo. Começa por recolher-se os dados, depois ordenam-se por ordem decrescente de frequência e classificam-se de 1 até n, sendo 1 aquele que tem maior frequência. Para fazer o gráfico em escala logarítmica da frequência (eixo dos xx) vs a classificação (eixo dos yy), normalizam-se os dados. A normalização consiste na divisão de todos os valores no eixo dos xx pelo máximo que tomam, analogamente para o eixo dos yy. Por último, aproxima-se uma reta ao conjunto de dados do gráfico, no sentido dos mínimos quadrados (a reta é calculada minimizando a soma do quadrado dos resíduos).

## 3.1. NÚMERO DE VEZES QUE UMA PÁGINA DA INTERNET É ACEDIDA

O objetivo deste estudo, sobre o número de vezes que uma página da internet é acedida, consiste em verificar se os acessos aos sítios de jogos e músicas seguem uma distribuição do tipo LP. Os dados obtidos para este estudo foram retirados do sítio: http://top.ucoz.com, em Abril de 2012.

Para cada caso foi construído um gráfico do tipo log-log em função da frequência/rank. Os dados obtidos foram tratados e processados. Foi possível concluir que, no caso da música, o número de acessos variou entre 1 e 214435. Observou-se, assim que existem muitos acessos às páginas com este tipo de conteúdos, conforme se pode observar na Figura 1. Os dados foram convertidos para valores relativos e apresentados num gráfico com escalas logarítmicas em ambos os eixos.

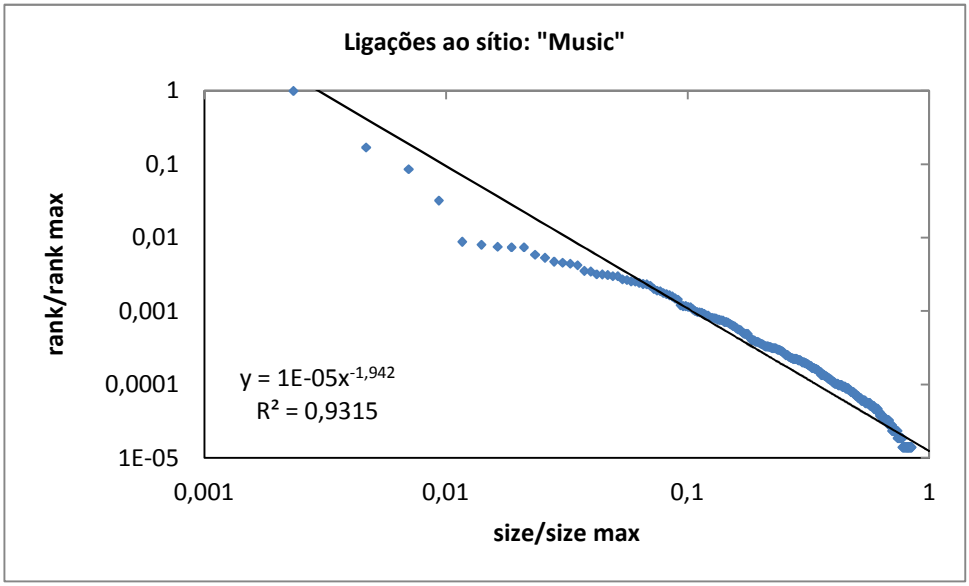



Figura 1 **Número de acessos registados pelo sítio: "*Music*".**

Outro tema bastante interessante, com muitos dados a serem processados e analisados, foi o sítio com o conteúdo de jogos, onde foram registadas variações no número de acessos entre 1 e 39385. Foi contabilizado um elevado número de acessos às páginas com os conteúdos relacionados com conteúdos de jogos: online, computador e para consolas. Os dados foram convertidos para valores relativos e apresentados num gráfico que segue uma LP, conforme indicado na Figura 2.

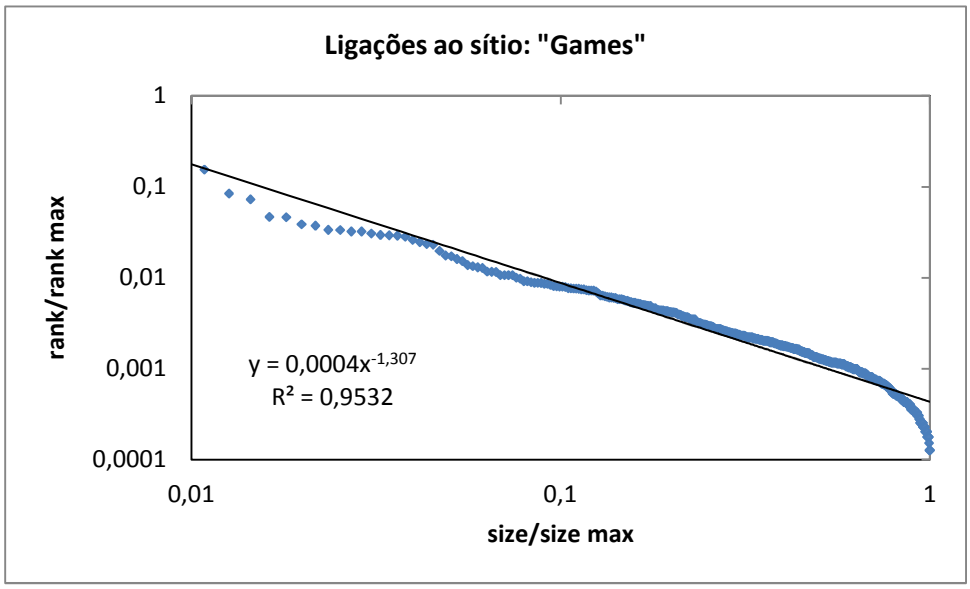



Figura 2 **Número de acessos registados pelo sítio: "*Games*".**

O sítio com conteúdos relacionados com *software* obteve um número de acessos compreendidos entre 1 e 13108, como se pode observar na Figura 3.

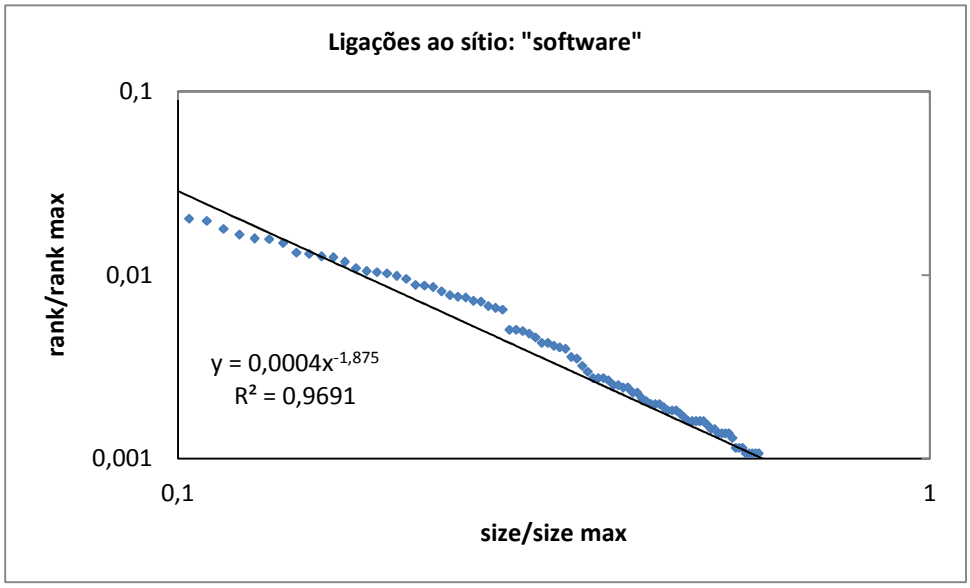



Figura 3 **Número de acessos registados pelo sítio: "*Software*".**

Grande parte da população mundial tem preferência por algum género de filme (*movies)*, assim sendo o sítio com conteúdos relacionados com filmes registou o número de acessos compreendidos entre 1 e 104720, como se pode observar na Figura 4.

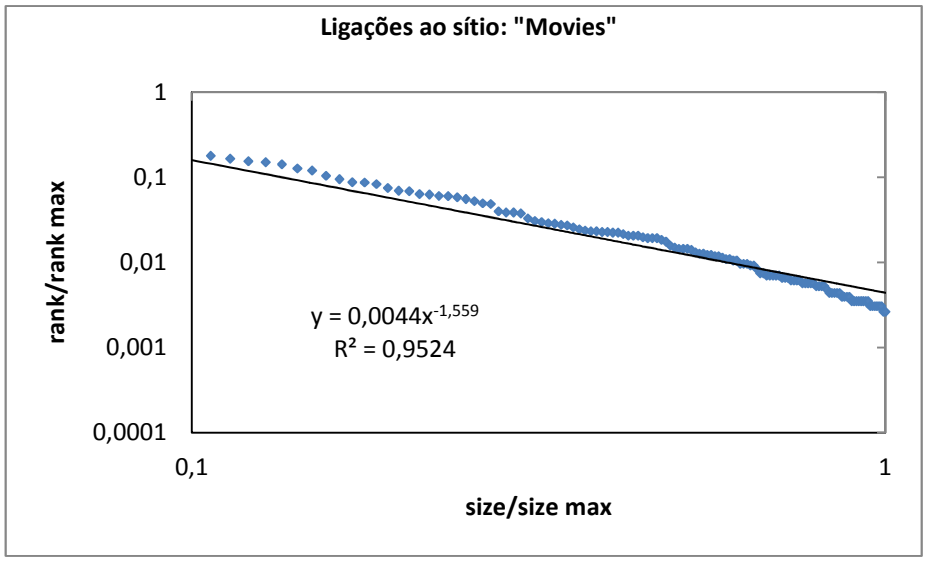



Figura 4 **Número de acessos registados pelo sítio: "*Movies*".**

O sítio com conteúdos relacionados com as diversas modalidades e informações sobre o mundo desportivo (*sports)*, registou um número de acessos compreendidos entre 1 e 2217, como se pode observar na Figura 5.

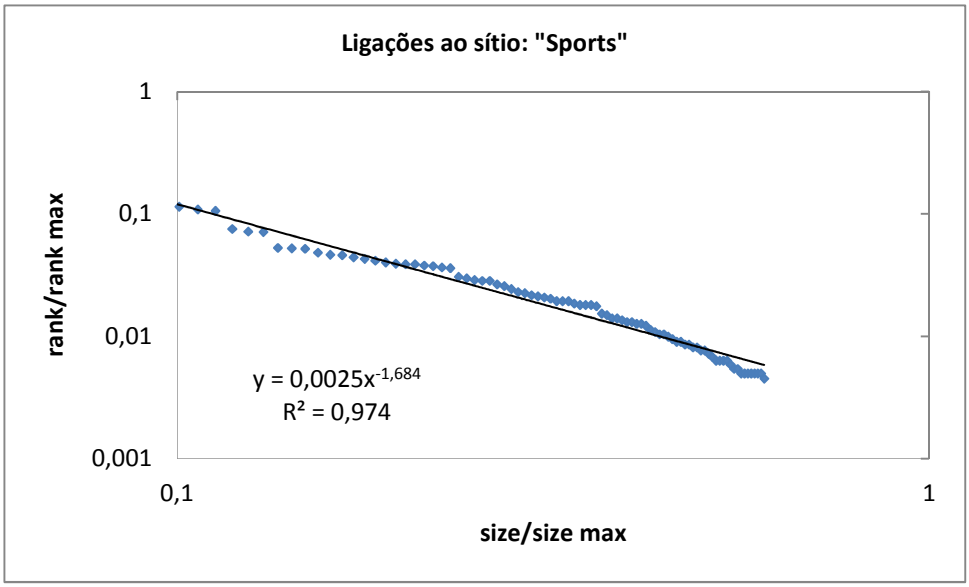



Figura 5 **Número de acessos registados pelo sítio: "Sports".**

Para o caso do sítio com conteúdos relacionados com educação foi registado um número de acessos compreendidos entre 1 e 31117, como se pode observar na Figura 6.

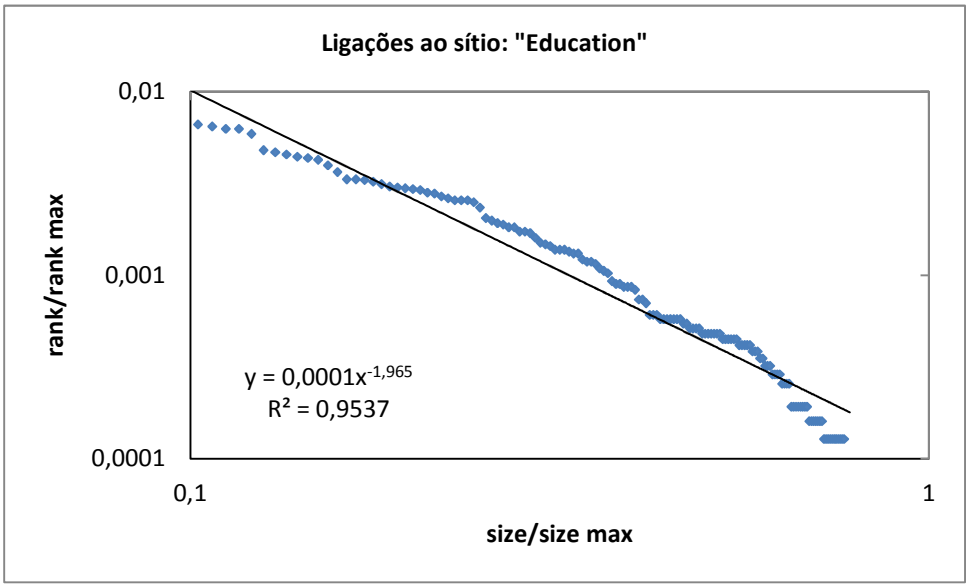



Figura 6 **Número de acessos registados pelo sítio: "*Education*".**

## 3.2. NÚMERO DE MÁQUINAS CONECTADAS A UMA REDE/PÁGINA

O objetivo deste estudo, sobre o número de máquinas conectadas a uma página, consiste em verificar se as ligações para dentro e fora dessa página seguem uma distribuição do tipo LP.

Consoante o interesse manifestado por um utilizador em encontrar determinado conteúdo na Web, é necessário estabelecer a ligação da sua máquina a uma página desejada. Aproveitando esta necessidade foram recolhidos e processados os dados associados a esta atividade. O número de máquinas ligadas aos sítios com conteúdos relacionados com a música variou entre 1 e 2905. Os dados foram trabalhados e obteve-se uma distribuição de LP, como se pode observar na Figura 7.

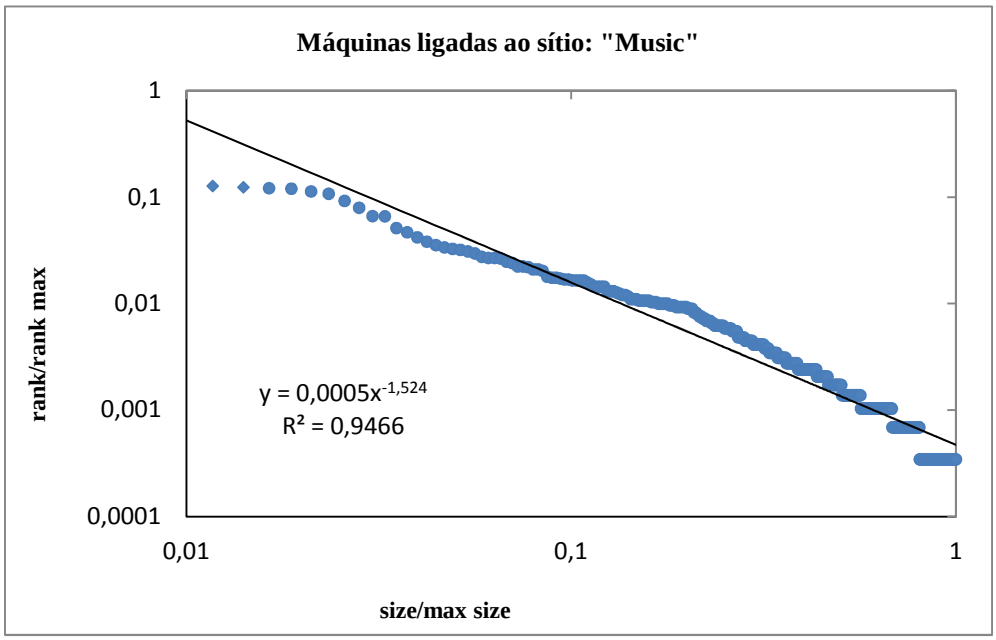



Figura 7 **Número de máquinas ligadas ao sítio: "Music".**

Associado ao crescimento, desenvolvimento e à necessidade de uma grande faixa da sociedade encontrar diversão, os sítios de jogos têm registado um elevado crescimento no número de máquinas ligadas. O valor do registo do número de máquinas que acedem ao sítio de jogos apresentou uma variação entre valor 1 e 2796. Na Figura 8 pode observar-se a existência de uma LP no crescimento do número de máquinas ligadas aos sítios de jogos na *Web*.

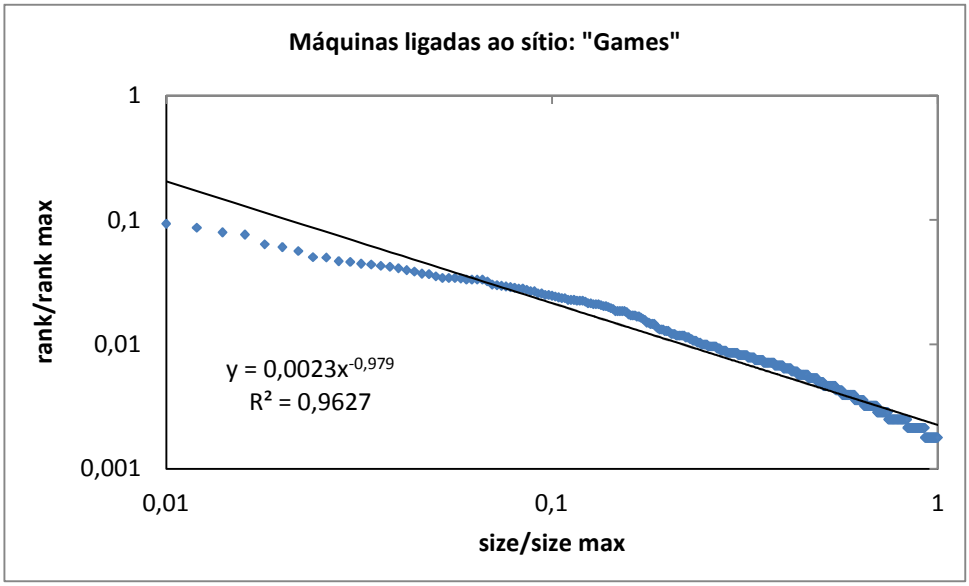



Figura 8 **Número de máquinas ligadas ao sítio: "*Games*".**

O crescimento da *Web* tem proporcionado uma maior publicação de conteúdos, o que leva um maior número de utilizadores a pesquisarem conteúdos relacionados com programação para conseguirem o melhor rendimento das suas máquinas. Foi registado um número compreendido entre 1 e 765 máquinas que estavam ligadas ao sítio *Software*, como é possível observar na Figura 9.

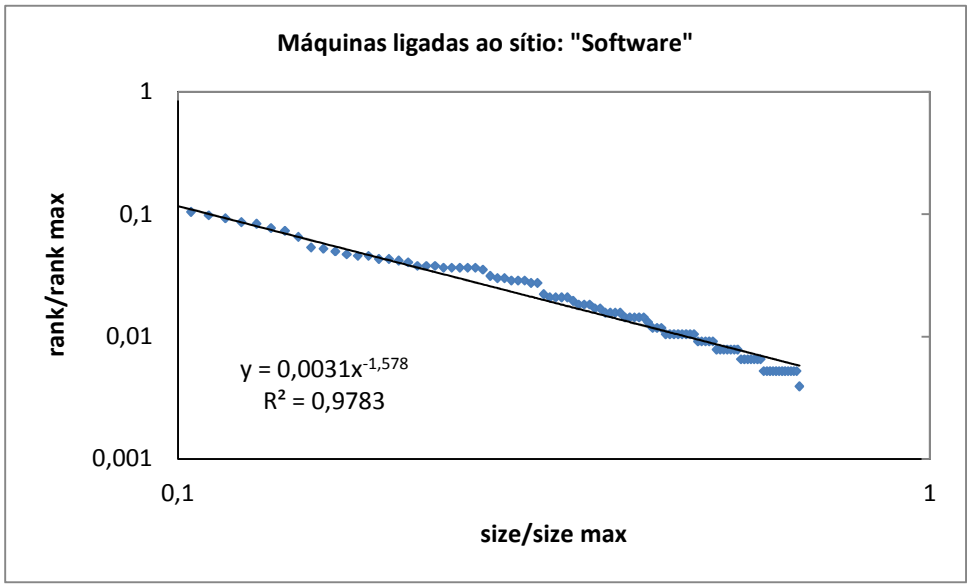



Figura 9 **Número de máquinas ligadas ao sítio: *"Software"*.**

O sítio com conteúdos relacionado com filmes (*Movies)* registou um número compreendido entre 1 e 9398 máquinas ligadas, como é possível observar na Figura 10.

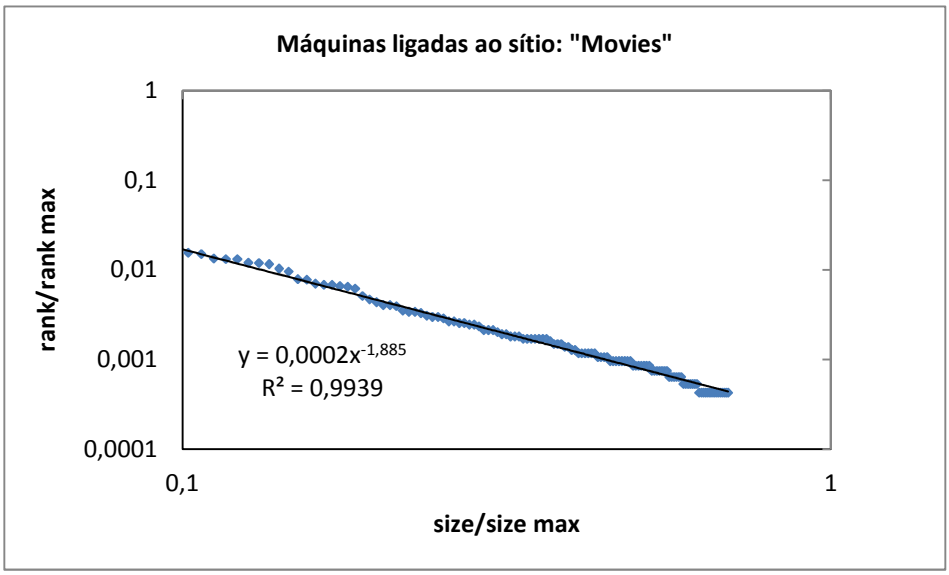



Figura 10 **Número de máquinas ligadas ao sítio: "*Movies*"**

Para o caso relacionado com desportos foi registado um número compreendido entre 1 e 1040 máquinas que procuravam informação ou realizam atividades deste interesse, como é possível observar na Figura 11.

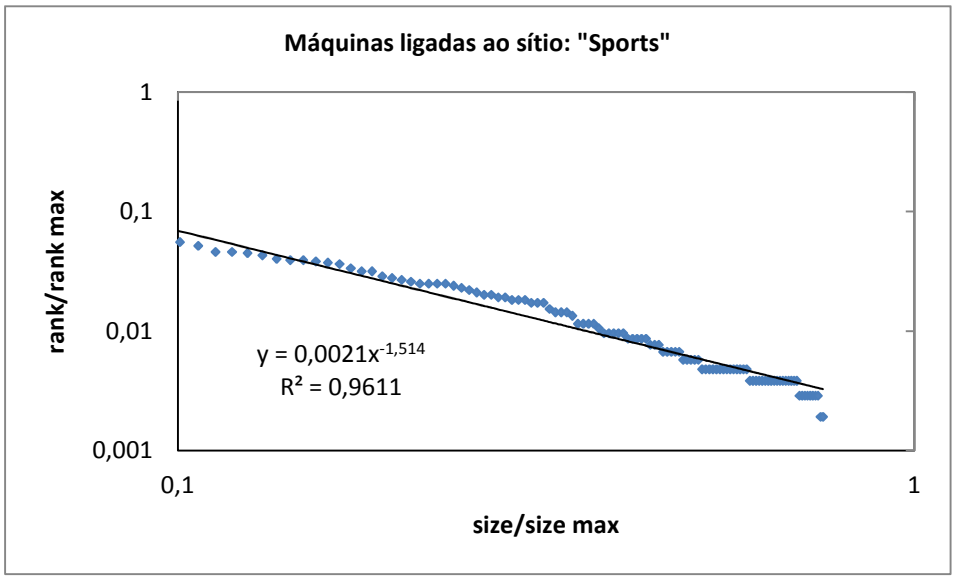



Figura 11 **Número de máquinas ligadas ao sítio: "*Sports*".**

O sítio com o conteúdo relacionado com educação apresentou um número compreendido entre 1 e 398 de máquinas ligadas, o que leva a concluir que não representa grande interesse por parte dos utilizadores como é possível observar na Figura 12.

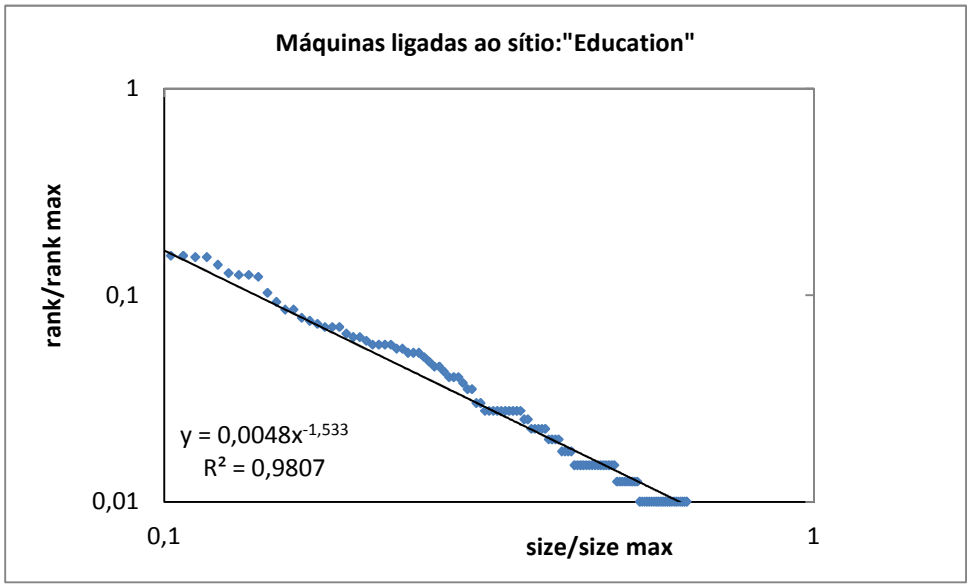



Figura 12 **Número de máquinas ligadas ao sítio:*"Education"*.**

## 3.3. DESEMPENHO DE COMPUTADORES

Desde 1993 o sítio Top500Project, http://www.top500.org tem como objetivo acompanhar o desenvolvimento e a evolução dos computadores, com um desempenho considerado acima da média, disponíveis no mercado. Semestralmente é editado, no seu sítio, o *rank* dos melhores computadores e o seu respetivo fabricante.

O desempenho de cada computador é avaliado segundo o parâmetro *Linpack*, que é testado a partir da resolução de equações lineares do tipo: A = b * X utilizando uma matriz aleatória densa. Este parâmetro está associado à expansão do número de computadores nos últimos 15 anos e o mesmo tem fornecido valores de desempenho acima da média. Até ao momento não é possível encontrar melhor relação entre a eficácia Rmax/Rpeak de um sistema. Nos EUA este parâmetro já é considerado de referência quando se escolhe um computador.

Quando um fabricante coloca no mercado um novo computador, é necessário realizar o teste do seu desempenho máximo (Rmax). Este teste permite conferir se as especificações estabelecidas para o equipamento são atingidas nas condições mais adversas de funcionamento. A partir dos inúmeros testes realizados pelo sítio top500Project, observou-se uma variação no intervalo compreendido entre 1 e 10510. É possível observar a existência de uma LP no desempenho das máquinas testadas num determinado período de tempo como está indicado na Figura 13.

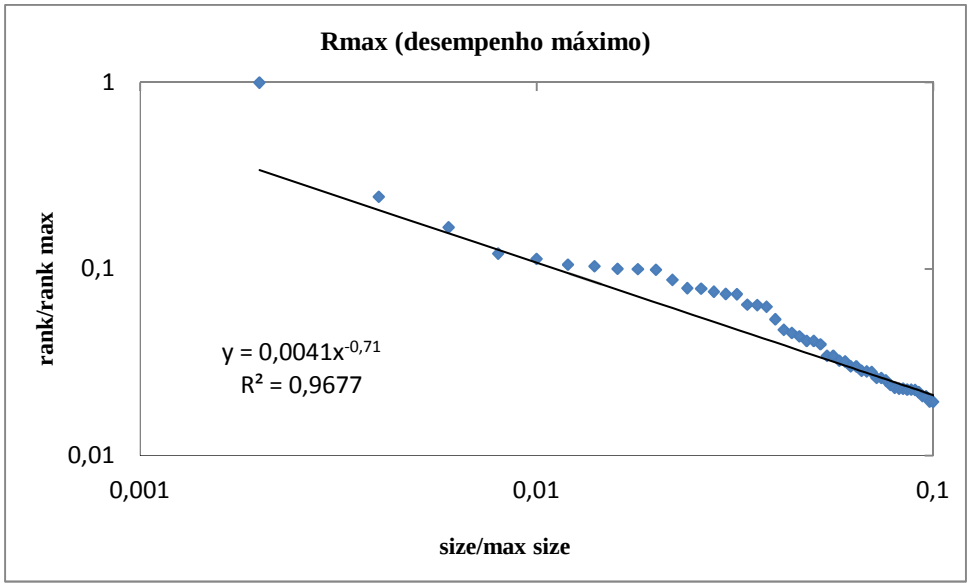



Figura 13 **Desempenho máximo alcançado por um computador.**

Nem sempre as especificações do equipamento são confirmadas depois da realização do teste Linpack assim, é atribuído o valor do desempenho teórico máximo (*Rpeak*), atingido por um determinado equipamento. Foram observadas uma quantidade de máquinas compreendidas no intervalo de 1 a 11280. Depois do registro e tratamento dos dados, foi construído o gráfico da Figura 14 onde é possível observar a LP no desempenho dos computadores testados.

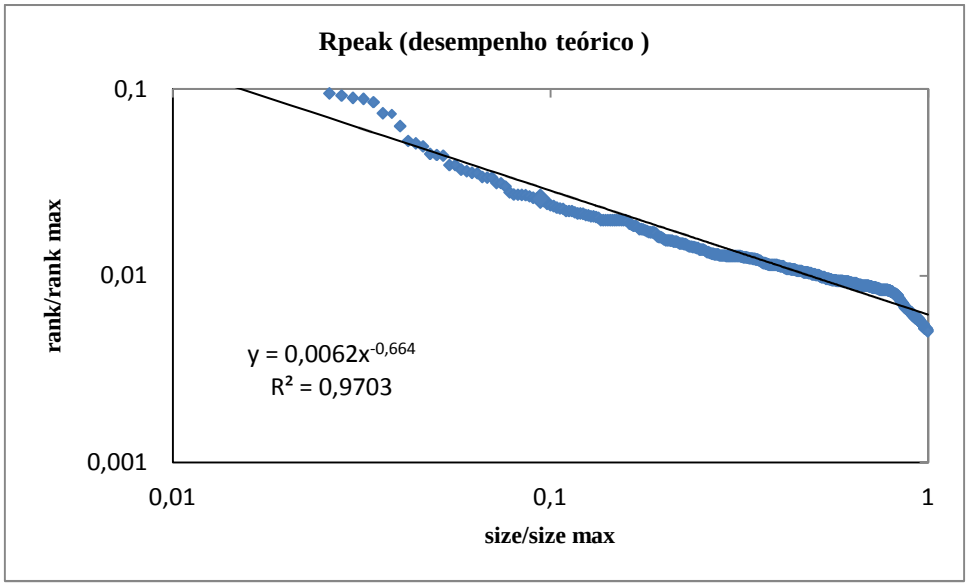



Figura 14 **Desempenho teórico máximo alcançado por um computador.**

## 3.4. NÚMERO DE LIGAÇÕES QUE APONTAM PARA DETERMINADO NÓ DE UMA REDE

A Web é constituída por diversos nós interligados onde circulam elevados fluxos de informações. Existem ligações realizadas de forma direta que permitem ligar dois ou mais nós e proporcionam maior rapidez na troca de dados. Na Figura 15 é possível observar a distribuição das ligações que apontam diretamente a um nó.

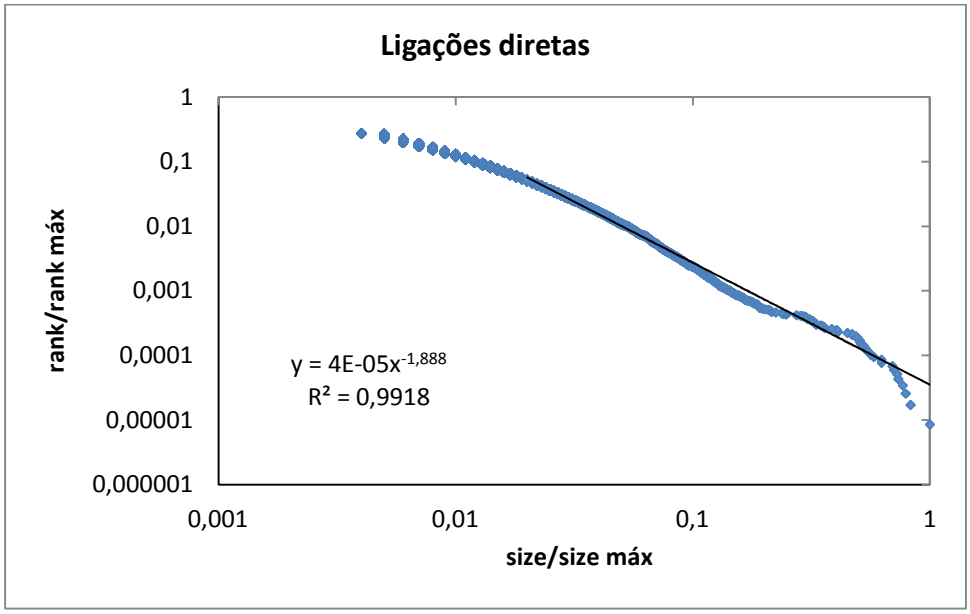



Figura 15 **Ligações que apontam diretamente a um nó.**

Existem situações onde as ligações não apontam diretamente ao nó, neste caso a ocorrência de congestionamento da rede apresenta maior probabilidade. Este fato vai influenciar a transferência de ficheiros e algumas operações realizadas por parte dos utilizadores. Na Figura 16 é possível observar a distribuição das ligações indiretas.

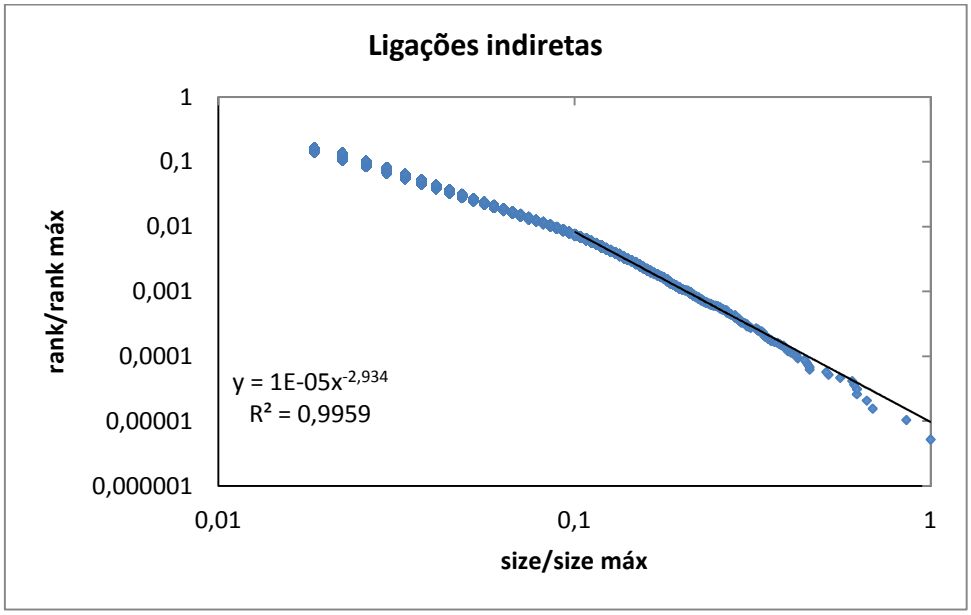



Figura 16 **Ligações que apontam para fora do nó.**

## 3.5. PERSPETIVA GLOBAL

Nesta subseção, foram apresentados vários exemplos de aplicações da Lei de Potência (LP) a fenómenos associados a internet. A Figura 17 apresenta o gráfico que combina a relação existente entre os parâmetros $\tilde{C}$ e $\alpha$ da distribuição de Pareto dos vários exemplos estudados. Observa-se que o desempenho de uma máquina (*Rank* ou *Rpeak*) apresenta valores de $\tilde{C}$ mais elevados em comparação com o número de ligações que aponta para um nó. O comportamento apresentado pela distribuição do número de *Hits* e *Hosts* foi aproximadamente constante para os valores de $\tilde{C}$ e $\alpha$ que estão muito próximos uns dos outros e esses valores estão mais dispersos no eixo do $\tilde{C}$. Esta caraterística revela uma distribuição constante para o número de ligações realizadas por utilizadores ou máquinas que acedem a um determinado sítio em busca de informações.

Em conclusão, a relação existente entre $(\tilde{C};\alpha)$ revela características globais dos fenómenos descritos pela LP e novas conclusões poderão ser retiradas a partir da adição de novos dados e fenómenos.

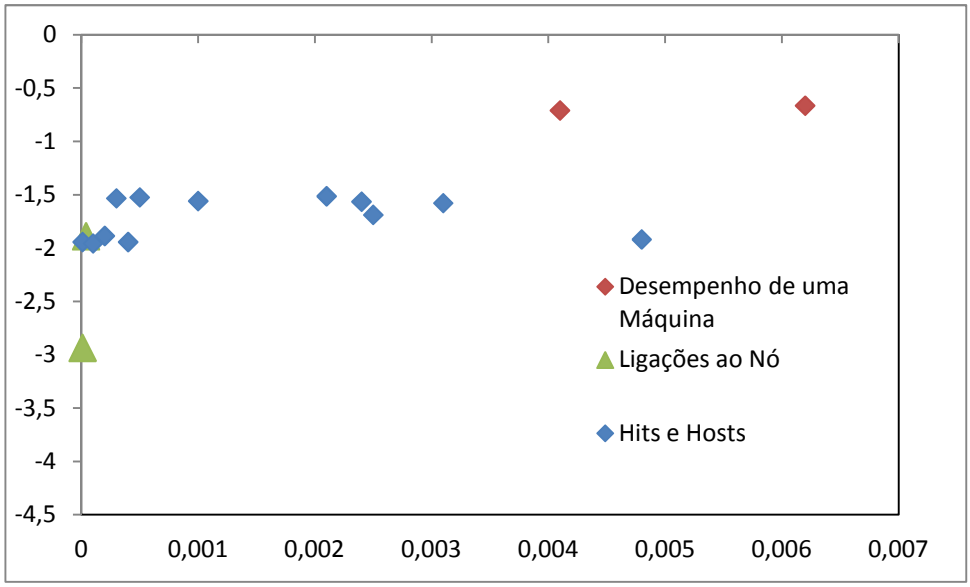



Figura 17 **Evolução do parâmetro $\tilde{C}$ em função de $\alpha$.**

# 4. CONCLUSÕES

Este trabalho tem como objetivo estudar as Leis de Potência (LP) e a sua aplicação aos fenómenos associados à internet. Por exemplo, uma LP modela o número de utilizadores que acedem um determinado sítio, distribuição de nós numa rede, número de livros vendidos, realização de leilões, propagação de vírus, ao número de *links* que constituem este sítio. Foram recolhidos dados reais disponíveis em sítios seguros e foram desenhados os gráficos respetivos.

Em escala logarítmica, em ambos os eixos, os dados são bem aproximados por uma reta, com declive negativo. Os resultados obtidos permitem-nos concluir que todos os dados estudados são bem aproximados, numa escala logarítmica, por uma reta com declive negativo, seguindo, assim, uma distribuição de Pareto.

O crescimento exponencial ao nível dos utilizadores, sítios e dos conteúdos encontrados na *Web,* condiciona a existência de muitos estudos. Alguns dos exemplos de aplicação das LP presentes neste trabalho, seguramente serão alvo de mais análise, observações e novas conclusões serão apresentadas. A importância da internet no nosso cotidiano é avassaladora e grandes investimentos continuarão a ser realizados, de forma a potenciar ainda mais essa importância. Este fato permite-nos concluir que ainda existe muito

trabalho em desenvolvimento ao nível teórico e prático sobre as Leis de Potência e sua aplicação à internet.

# *Referências Documentais*


[1] L.A. Adamic. The Small World Web, Proceedings of ECDL'99. Lecture Notes in Computer Science 1696, 443-452. Berlin: Springer (1999).

[2] L.A. Adamic, B.A. Huberman. Technical comment: power-law distribution of the world wide web. Science 2000; 287: 2115A.

[3] L. A. Adamic, R. M. Lukose, A. R. Puniyani and B. A. Huberman. Search in Power-Law Networks. Physical Review E 64: 046135 (2001).

[4] L.A. Adamic, B.A. Huberman. The Web's Hidden Order. Communications of the ACM. Volume 44, Issue 9, pag. 55-59 (2001).

[5] L.A. Adamic, B. A. Huberman. Zipf's Law and the Internet. Glottometrics 3, 143-150 (2002).

[6] L. A. Adamic and B. A. Huberman. The Web's Hidden Order Hewlett-Packard Labs Palo Alto, CA 94304.

[7] R. Albert, H. Jeong, and A. L. Barabasi. The Diameter of the World-Wide Web, Natura, 401:130, (1999).

[8] Y. Aljure & J. Gallego. Desigualdad y leyes de potencia. Cuadernos de Economía, 29(53) (2010).

[9] R. Albert, H. Jeong and A.-L Barabasi. Attack and error tolerance of complex networks. Nature 406, 378 (2000).

[10] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, pp. 171–180, Association of Computing Machinery, New York (2000).

[11] M. Batty. Cities and Complexity: Understanding Cities Through Cellular Automata, Agent-Based Models, and Fractals MIT Press, Cambridge, MA, (2005).

[12] L. Breslau *et al.* Web Caching and Zipf-like Distributions: Evidence and Implications. Proceedings of INFOCOM '99, 126-134 (1999).

[13] E. Brynjolfsson, Y.J. Hu, and M.D. Smith. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. MIT Sloan Working Paper No. 4305-03. Available at SSRN: http://ssrn.com/abstract=400940 or doi:10.2139/ssrn.400940 June (2003).

[14] J. P. Bouchard and M. Potters, Theory of Financial Risks: From Statistical Physics to Risk Management Cambridge University Press, Cambridge, (2000).

[15] L.E. Cederman. Modeling the Size of Wars: From Billiard Balls to Sandpiles. Amer. Polit. Sci. Rev. 97, 135–150 (2003).

[16] J. Chevalier, A. Goolsbee. Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. Quantitative Marketing and Economics 1 203–222, (2003).

[17] CERT Coordination Center, Carnegie Mellon University, http://www.cert.org/advisories/CA-2000-04.html.

[18] A. Clauset, M. Young, K.S. Gleditsch. On the Frequency of Severe Terrorist Events. Journal of Conflict Resolution 51 No 1 58–87, (2007).

[19] C. R. Cunha, A. Bestavros, and M. E. Crovella. Characteristics of WWW Clientbased Traces". Technical Report TR-95-010. Boston University Computer Science Department (1995).

[20] H. Ebel, L. I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E* 66, 035103 (2002).

[21] J. B. Estoup, Gammes Stenographiques. Institut Stenographique de France, Paris (1916).

[22] M. Faloutsos, P. Faloutsos and C. Faloutsos. On Power-Law Relationships of the Internet Topology. Proceedings of ACM SIGCOMM '99, 251-262 (1999).

[23] R. H. Frank and P. J. Cook, The Winner-take-all Society, Free Press, New York, NY (1995).

[24] X. Gabaix. Zipf's Law and the Growth of Cities. American Economic Review Papers and Proceedings LXXXIX 129–132 (1999).

[25] X. Gabaix. Zipf's law for cities: an explanation. The Quarterly Journal of Economics 114, 739–767 (1999).

[26] R. Gibrat. Les in'egalit'es 'economiques. Paris, France, Librairie du Recueil Sirey (1931).

[27] A. P. Hackett, 70 Years of Best Sellers. 1895-1965. R. R. Bowker Company, New York, NY (1967).

[28] B.A. Huberman, P.L.T. Pirolli, J.E. Pitkow, and R.M. Lukose. Strong regularities in world wide web surfing. Science, 280:95–97, (1998).

[29] B.A. Huberman and L.A. Adamic. Evolutionary Dynamics of the World Wide Web", Nature 401, 131. Nature (London) 401, 131 (1999).

[30] B. A. Huberman and L. A. Adamic. Information dynamics in the networked world. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai (eds.), Complex Networks, number 650 in Lecture Notes in Physics, pp. 371–398, Springer, Berlin (2004).

[31] H. Jeong, R. Albert and A.-L. Barabasi. Diameter of the World Wide Web. Nature 401, 130 (1999).

[32] P. Krugman. The Self-Organizing Economy, Cambridge, MA: Blackwell (1996).

[33] M J.G. Kemeny and J.L. Snell. Finite Markov Chains. D. Van Nostrand, Princeton, NJ, (1960).

[34] Levene and G. Loizou. Navigation in hypertext is easy only sometimes. SIAM Journal on Computing, 29:728–760, (1999).

[35] M. Levene and G. Loizou. Web interaction and the navigation problem in hyper-text. In A. Kent, J.G. Williams, and C.M. Hall, editors, Encyclopedia of Micro-computers. Marcel Dekker, New York, NY, To appear (2000).

[36] R.M. Lukose and B.A. Huberman. Surfing as a real option. In Proceedings of the International Conference on Information and Computation Economies, pages 45–51, Charleston, SC, (1998).

[37] R. N. Mantegna and H.E. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance Cambridge University Press, Cambridge, (2000).

[38] S. Miyazima, Y. Lee, T. Nagamine, and H. Miyajima, Power-law distribution of family names in Japanese societies. Physica A 278, 282–288 (2000).

[39] G. Neukum and B. A. Ivanov, Crater size distributions and impact probabilities on Earth from lunar, terrestialplanet, and asteroid cratering data. In T. Gehrels (ed.), Hazards Due to Comets and Asteroids, pp. 359–416, University of Arizona Press, Tucson, AZ (1994).

[40] M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323–351 (2005).

[41] P. Satarros, R. and A. Vespignani. Epidemic spreading in Scale Free Networks. Physical Review Letters 86, 3200 (2005).

[42] D. J. de S. Price, Networks of scientific papers. Science 149, 510–515 (1965).

[43] L.F. Richardson. Variation of the Frequency of Fatal Quarrels with Magnitude. J. Amer. Statistical Assoc. 43, 523–546 (1948).

[44] L.F. Richardson. Statistics of Deadly Quarrels. Chicago: Quadrangle Books (1960).

[45] E. Van Heck, O. R. Koppius, & P. H. Vervest. Electronic web-based auctions: theory and practice. In W.R.J. Baets (Ed.), Proceedings 6th European Conference on Information Systems, Volume IV (pp. 1584-1590). Granada, Spain: Euro-Arab Management School (1998).

[46] I. Yang, H. Jeong, B. Kahng, and A.-L. Barabási. Emerging behavior in electronic bidding, Physical Review E 68 (2003).

[47] G. Zipf. Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press, cambridge, MA (1932).

[48] G. K. Zipf, Human Behaviour and the Principle of Least Effort. Addison-Wesley, Reading, MA (1949).