Une analyse descriptive de la fracture numérique des usages en ligne à partir des données de navigation

Fabrice Le Guel
CREM – UMR CNRS 6211

ICTs and inequalities : the digital divides
TIC et inégalités : les fractures numériques
International conférence / Colloque international
18 – 19 novembre 2004
Carré des Sciences, rue Descartes, Paris, France

Résumé :

Dans le prolongement des études relatives à la fracture numérique de 'premier niveau' (celle de l'accès à Internet), une récente littérature a pointé sur l'existence d'une fracture numérique de 'second niveau' (celle des usages en ligne), qui décrit la capacité des internautes à utiliser Internet de façon « effective et efficiente ».

Parallèlement à cette première littérature, un autre thème de recherche s'est penché sur la mesure des usages en ligne à partir des données de navigation. Le principal résultat de cette seconde littérature est qu'il existe une loi de navigation ou peu de sites reçoivent la majorité des visites.

Nous proposons dans cette contribution de faire le lien entre ces deux littératures. Nous étudions dans quelle mesure il est possible d'utiliser des données de navigation pour évaluer la capacité des individus à utiliser Internet. Nous montrons que l'analyse des comportements de navigation doit se faire non pas au niveau agrégé (les sites), mais plutôt au niveau désagrégé (les internautes). Nous discutons enfin du rapport entre l'hétérogénéité des comportements de navigation et l'inégalité des comportements de navigation.

Abstract :

In the continuation of the studies relating to 'first-level' digital divide (the Internet access), a recent literature pointed on the existence of a second-level digital divide (the internet use) which describes differences in people's online skills. Parallel to this first literature, another research topic considered the measurement of the Internet uses using 'clickstream data'. The main result of this second literature is that there is a 'surfing law' that most sites get only a few visits.

In this paper, we suggest to establish the link between these two literatures. We study up to what point it is possible to use clickstream data to measure the online ability. We show that the analysis of the Internet behaviors must be done not at an aggregate level (the Web sites), but rather at the disaggregate level (the Internet users). Finally, we discuss the relationship between the heterogeneity and inequality of online behaviors.

Mots clés : Internet, Usages, Power Law, Zipf law, Mesure, Comportements de navigation, Clickstream Data, Fracture numérique de second niveau.

1. Introduction :

Dans le prolongement des études relatives à la fracture numérique de 'premier niveau' (celle de l'accès à Internet), une récente littérature a pointé sur l'existence d'une fracture numérique de 'second niveau', celle des usages en ligne (Hargittai, 2002, DiMaggio et al., 2004). Au sens d'Hargittai (2002), cette seconde fracture sépare les internautes qui sont aptes à utiliser Internet « de façon effective et efficiente », de ceux qui n'en sont pas capables.
Parallèlement à cette première littérature, un autre thème de recherche s'est penché sur la mesure des usages en ligne à partir des données de navigation (Goldfarb, 2002a, 2002b ; Montgomery & al., 2004). Souvent appelées données Log ou encore *clickstream data*, les données de navigation permettent, sous certaines conditions techniques, de relever l'ensemble des comportements de visites ou de clics sur les sites Internet visités par un échantillon d'utilisateurs.
Nous proposons dans cette contribution de faire le lien entre ces deux pans de recherche et d'utiliser alors les données de navigation pour évaluer, au sens d'Hargittai (2002), la capacité des internautes à naviguer sur différents sites. Par ce biais, nous posons d'abord la question de la mesure de usages en ligne, essentielle si nous voulons vérifier d'une manière quantitative l'existence d'une fracture de 'second niveau'. Mais la question principale de cet article vient d'un constat issu de la littérature chargée de mesurer les usages sur Internet à partir des données de navigation. Cette littérature montre en effet qu'il existe une loi de navigation - ou *surfing law* - (dite 'universelle' – Adamic & Huberman, 2000, 2001, 2002) où la majorité des sites reçoivent une minorité de visites et réciproquement, une minorité de sites reçoivent la majorité des visites. Notons que cette loi de navigation est correctement restituée par une distribution asymétrique telle que la loi de Zipf (*Zipf law*), la loi de Pareto ou encore la loi puissance (*power law*). Dès lors, faut-il voir dans l'existence d'une telle loi une homogénéité des comportements de navigation en ligne, auquel cas il n'y aurait pas véritablement de différences dans la capacité des individus à utiliser Internet ?
Pour répondre à cette question, nous avons besoin de travailler sur des données de navigation. Nous proposons alors d'utiliser un échantillon de 584 étudiants dont les activités de navigation ont été observées durant trois mois en 1995. Notre approche sera donc descriptive. Nous montrons que les *surfing law* permettent de restituer uniquement une fracture numérique (éventuelle) du côté de l'offre (les sites Internet) et non pas de la demande (les internautes). Pour cela, nous transitons d'une analyse 'agrégée' des comportements de navigation (qui se concentre sur les sites Internet) vers une analyse 'désagrégée' des comportements de navigation (qui se situe au niveau des visites de chaque internaute). Nous posons alors la question de l'existence d'une loi de navigation *au niveau individuel* et nous nous interrogeons sur la conséquence d'une telle loi du point de vue de la reconnaissance d'une fracture numérique de second niveau.
Cet article s'organise comme suit. Nous présentons dans la section 3 un nouveau format de données pour les Sciences Humaines - les données de navigation - ainsi que la littérature ayant utilisé ces dernières. La section 4 se charge de construire (à l'aide de notre échantillon d'étudiants) les distributions asymétriques observées dans la littérature. La section 5 propose de dépasser les résultats précédents en construisant la distribution des lois puissances pour chaque internaute. Nous tirons les conséquences de ces distributions du point de vue de la fracture des usages en ligne.
Avant d'arriver à cela, la section suivante (section 2) présente la notion de fracture numérique de 'second niveau'.

2. La 'double fracture numérique'

2.1 La fracture de 'premier niveau' : celle de l'accès à Internet

La notion de 'fracture numérique' est semble-t-il pour la première fois apparue en 1995 , aux Etats-Unis, sous l'expression '*digital divide* '. Néanmoins, il est impossible d'en attribuer la paternité. Par exemple, selon Compaine (2001) , L. Morrisset, alors président de la fondation Markle (chargée de stimuler l'adoption des TIC dans le domaine de la santé et de la sécurité), aurait le premier prononcé ces mots. Plusieurs journalistes pourraient aussi être à l'origine de ce terme, par exemple J. Webber et A. Harman, du Los Angeles Times, G. A. Poole, du New York Times, ou encore Long-Scott, du journal Outlook.
Il est encore plus difficile de s'accorder sur une définition unifiée de la notion de fracture numérique. Cette dernière est souvent « vague et extensive (des questions d'infrastructures de télécommunication aux programmes d'éducation) s'appliquant à des situations très différentes (nations, régions, organisations, communautés, groupes sociaux, individus…) » (Rallet & Rochelandet, 2003, p. 2). Pour s'en rendre compte, nous pouvons examiner la définition proposée par l'OCDE (2001) : « *the term 'digital divide' refers to the gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard both to their opportunities to access information and communication technologies (ICTs) and to their use of the Internet for a wide variety of activities. The digital divide reflects various differences among and within countries* » (ibid, p. 4).
La fracture numérique concerne selon l'OCDE l'ensemble des biens issus des TIC. En effet, plusieurs fractures peuvent exister, aussi bien du point de vue de l'objet (les téléphones portables, l'Internet ; Rice & Katz, 2003) que du sujet (les ménages, les entreprises ; Forman, 2003). Montagnier, Muller, Vickery (2002) proposent de réduire le champ d'analyse à l'Internet et un de ses supports, les outils informatiques : « *differences in access to information and communication technologies (ICTs), such as computers and the Internet, create a "digital divide" between those that can benefit from opportunities provided by ICTs and those that cannot* » (ibid, p.1). Cette vision 'binaire' de la fracture numérique est la plus commune dans la littérature. Elle oppose les individus qui ont accès aux TIC et particulièrement à Internet, donc au 'savoir' (les '*information haves*'), à ceux qui n'ont pas accès à cette information, (les '*information have nots*').
Néanmoins, même cette approche restreinte de la fracture numérique peut porter à confusion, du fait d'une perception relativement fluctuante de la notion d'accès à l'Internet. DiMaggio et al. (2004) montrent par exemple que l'importance de la fracture numérique varie en fonction de ce que l'on entend par 'accès à Internet'. Ce dernier peut être plus ou moins conditionné au lieu dans lequel on se trouve (au travail, à domicile, chez des amis, dans un cyber-café ou à partir d'un Point d'Accès Public à Internet - PAP). On peut alors parler d'une fracture numérique, tous lieux confondus, ou se concentrer uniquement sur les inégalités dans l'accès des ménages ou encore sur le lieu de travail…
Il existe donc de multiples façons d'envisager la notion de fracture numérique, à l'origine d'une littérature volumineuse (Rallet & Rochelandet, 2003 ; DiMaggio et al., 2004).

2.2 La fracture de 'second niveau' : celle des usages en ligne

Au-delà de la vision restrictive d'une fracture numérique qui ne concerne que ceux qui ont accès à Internet versus ceux n'y ont pas accès, quelques définitions plus larges ont été proposées par la littérature, en prenant désormais en compte les différences *parmi les internautes*, c'est-à-dire ceux qui accèdent déjà à Internet.
L'American Library Association (2003) propose par exemple de définir la fracture numérique de cette façon : « *Differences due to geography, race, economic status, gender, and physical ability in [1] access to information through the Internet, and other information technologies and services, [2] in the skills, knowledge, and abilities to use information, the Internet and other technologies* » . Nous avons dès lors une définition de la

fracture à deux niveaux, le premier niveau concerne les inégalités d'accès, le second s'intéresse aux inégalités d'usages, désignées par les auteurs comme les différences dans l'aptitude à utiliser Internet, encore appelées « *skills gap* » .

Dès les premières études empiriques relatives à la fracture dans l'adoption d'une connexion Internet, plusieurs constats d'une fracture dans les usages apparaissent. En 1995, le rapport NTIA observe par exemple une inégalité dans la capacité des internautes à effectuer certaines tâches, notamment lorsqu'il s'agit de chercher des petites annonces sur Internet, des rapports gouvernementaux, ou de se former en ligne. De la même manière, Anderson et al. (1995) identifient des inégalités dans l'accès au service de courrier électronique. Néanmoins, ce phénomène n'est pas clairement explicité dans les précédentes études. Il faudra attendre Kling (1998) pour proposer une acception en diptyque de la fracture numérique. L'auteur identifie ainsi les inégalités dans l'accès aux TIC (il appelle cela *technical access*), des inégalités en termes de connaissances et de compétences techniques nécessaires pour bénéficier des TIC (appelé *social access*). Dans une approche voisine, Guichard (2003) et George (2004) parlent de l'accès à Internet mais aussi de son appropriation. Enfin, toujours dans la même logique, Le Guel, Pénard, Suire (2004, 2005), Le Guel (2004), développent l'idée d'une 'double fracture numérique', celle de l'accès et celle de l'usage en ligne.

A notre connaissance, Hargittai (2002, 2003) est le premier auteur à s'être réellement penché sur les causes possibles d'un tel phénomène. Selon l'auteur, la fracture de second niveau est une fracture cognitive, séparant les internautes qui ont la « capacité à trouver de façon effective et efficiente des informations en ligne » (ibid., 2002, p.2) de ceux qui ne l'ont pas. L'efficience est jugée en fonction de deux critères. Le premier indique si la tâche demandée a été effectuée (il s'agit par exemple de trouver des informations politiques ou culturelles en naviguant sur Internet). Le second critère prend en compte le temps pour réaliser pleinement cette tâche. DiMaggio & Hargittai (2001), Hargittai (2002, 2003), DiMaggio et al. (2004) identifient alors cinq causes principales pouvant expliquer ces disparités dans la capacité à trouver de l'information sur Internet :

1. l'inégalité des moyens techniques (type d'ordinateur utilisé, de logiciel et vitesse de connexion pour accéder à l'Internet),
2. les disparités dans l'autonomie d'utilisation de l'Internet (les individus accèdent-ils à Internet dans plusieurs endroits, ont-ils besoin systématiquement de l'aide d'une tierce personne, sont-ils limités en temps de navigation parce qu'il faut partager la connexion ?),
3. la disparité des objectifs pour lesquels les individus utilisent Internet (rencontrer d'autres internautes, se former en ligne, trouver du travail, participer à la démocratie locale),
4. l'inégalité des compétences (pour utiliser un moteur de recherche, ou en terme de capacité à régler les problèmes techniques soi-même),
5. les différences dans le soutien social (permettant de faciliter l'utilisation d'Internet). Trois niveaux d'assistance sont proposés. Le premier concerne les individus payés pour fournir cette aide (au bureau ou via un service commercial), le second correspond à l'assistance de la famille, le troisième niveau d'aide concerne les échanges entre les individus ayant des intérêts et des compétences très proches.

Pour le moment, l'ensemble de ces sources possibles qui pourraient expliquer une fracture numérique de second niveau restent hypothétiques. La démarche d'Hargittai (2002) n'a fait qu'identifier des difficultés pour certains individus à effectuer des tâches déterminées sur Internet (une tâche pouvait par exemple consister à trouver des informations en ligne à propos des évènements culturels locaux). Pour cela, l'auteur a proposé d'observer (par enregistrements vidéo) les comportements de navigation d'un échantillon aléatoire de 100 internautes. Cette technique d'observation des usages en ligne reste néanmoins lourde. Après l'expérience (réalisée dans un laboratoire constitué d'ordinateurs), il faut visualiser les vidéos des navigations de chaque internaute et relever quels sites Internet ont été visités, le temps passé sur chacun d'entre eux,

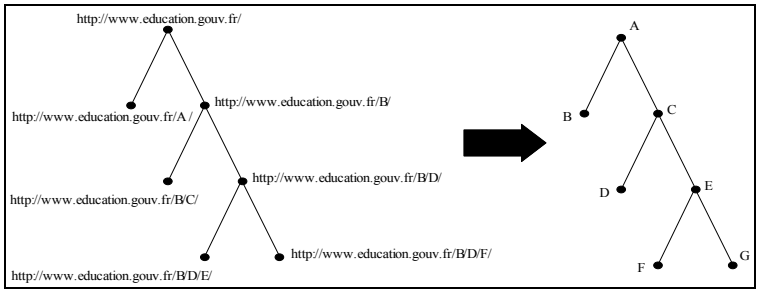
puis vérifier si la tâche demandée a abouti. En réalité, l'expérience d'Hargittai a montré que l'analyse des usages en ligne pose le problème de la mesure de ces derniers. En réponse à ces difficultés, DiMaggio, Hargittai et al. (2004) ont récemment proposé d'utiliser un format de données peu habituel en Sciences Humaines, à savoir les données (flux) de navigation, souvent appelées dans la littérature anglo-saxonne '*clickstream data*' : *A longterm priority is to go beyond self-reports by exploiting "clickstream" data detailed records (collected by market researchers) of the sites that individual Web users visit* [...] *Employing clickstream data presents many challenges --- gathering information about respondents' demographic traits and social attitudes without violating their privacy, classifying sites by topical domain, providing functional codes (e.g., shopping, playing games, gathering information) for particular visits based on information about the pages accessed, deducing when multiple users are employing the same account –-- that will require collaboration between social scientists and computer scientists. At the same time, because clickstream data are both behavioral and extremely detailed, they can answer questions (for example, what kinds of users access the highest quality information, or the extent to which users avoid or seek out sites that challenge their political views or aesthetic preferences) that survey data can only begin to address (Ibid, pp. 37-38).*
L'utilisation des données de navigation dans l'analyse des usages sur Internet fait partie d'un programme de recherche de long terme pour les auteurs précédents qui appartiennent à la sociologie. En économie et en marketing, un programme de recherche intitulé '*Choice and the Internet : from clickstream to research stream'* a été lancé aux Etats-Unis en 2002 (Bucklin et al., 2002). En France, le CNRS a proposé en 2003 une Journée d'Action Spécifique sur le thème de la 'mesure des flux Internet' (Lebart & Beaudouin). Cette action a été reconduite en septembre 2004. Enfin, un ouvrage rassemblant une large communauté de chercheurs français et canadiens a récemment traité de la mesure de l'Internet et des usages en ligne (Guichard, 2004). Le problème de la mesure des usages est donc d'actualité et la question de la reconnaissance d'une fracture de second niveau peut selon nous relever de cette problématique. Nous proposons donc dans les sections suivantes de travailler sur des données de navigation. Nous définirons en premier lieu ce format de données.

3. Les données de navigation et les résultats des analyses descriptives

3.1 Qu'est-ce que les données de navigation ?

Au-delà des enquêtes conventionnelles, depuis 1998/1999, un nouveau moyen d'analyser les usages en ligne s'est présenté. Les caractéristiques techniques du réseau Internet permettent en effet de rassembler une source d'informations, appelée 'fichiers Log' (Log files) par les informaticiens, et souvent renommée dans d'autres disciplines 'données de parcours' (Beaudouin & Licoppe, 2002), 'flux de données' (Lebart & Beaudouin, 2003), 'traces d'usages' (Montgomery & Faloutsos, 2000) ou encore '*clickstream data'* dans la littérature anglo-saxonne (Goldfarb, 2002a, 2002b, 2003). De notre côté, nous emploierons la notion générique de 'données de navigation'. Il existe deux sources de données de navigation, en fonction du lieu 'physique' où sont enregistrées ces informations. Nous parlons alors des données de navigation intra-site et des données de navigation inter-sites. Nous allons tout d'abord nous intéresser à la première source d'informations : les données de navigation intra-site.
Pour mieux identifier le format des données de navigation intra-site, nous proposons de présenter la structure simplifiée d'un site Internet (graphique 1).

Graphique 1: Exemple d'une structure simple de site Internet



Dans ce schéma, chaque point (ou nœud) correspond à une page Internet, et les segments reliant ces points annoncent qu'il existe un lien hypertexte amenant aux sous-branches immédiates de l'arbre. Notons qu'à titre illustratif, cette structure est simplifiée à l'extrême. Avec le développement de certains langages (le PHP par exemple) et l'utilisation des bases de données, la structure des sites est en effet devenue plus complexe : ces derniers sont rendus dynamiques, c'est à dire que leur architecture s'adapte en temps réel à la demande ou au profil des visiteurs. Cela complique le traitement des données de navigation, sans toutefois remettre en cause leur richesse potentielle.

Dans cette structure simplifiée, chaque page est matérialisée par une adresse Internet spécifique, appelée Uniform Resource Locator – URL, par exemple http://www.education.gouv.fr ou encore http://www.internet.gouv.fr. Cette adresse est souvent visible non loin de la barre de tâches des navigateurs Internet[1] et contient au moins trois types d'informations :

1. Tout d'abord, l'adresse URL permet d'identifier le nom du site (le nom de domaine). Nous parlons alors d'adresse 'racine', puisqu'elle correspond à la racine de l'arbre présenté dans le graphique 1 (exemple : http://www.education.gouv.fr). Rappelons d'autre part que l'extrémité de chaque adresse URL (dans l'exemple précédent, le '.fr' peut donner une première indication sur le contenu du site (site éducatif, commercial, institutionnel, associatif, etc) ;
2. ensuite, puisque la plupart des sites Internet possèdent plus d'une page, chaque page supplémentaire est identifiée par une adresse URL spécifique (exemple : http://www.education.gouv.fr/B/). Le nom de chacune des pages (B dans l'exemple précédent) procure donc des informations supplémentaires sur l'offre de biens ou services proposée par le site, il suffit pour cela d'identifier le contenu respectif des pages, en visionnant ces dernières sur Internet[2] ;
3. enfin, toujours à partir des adresses URL, il est parfois possible de caractériser certains usages, via les protocoles d'échange de l'information sur Internet. Chaque protocole est identifié au niveau de l'adresse racine. Le protocole 'http' est considéré comme générique (s'agissant de notre exemple : http://www.education.gouv.fr). Par contre, le protocole 'https' annonce que les informations transmises en ligne sont sécurisées. Il peut donc s'agir d'un achat en ligne (ou d'une volonté d'acheter). Le protocole 'ftp' (par exemple ftp://ncstrl-ftp.mit.edu) décrit un téléchargement ou un envoi d'informations en ligne, etc[3]. Dans certains cas, la simple lecture d'une adresse URL permet de rassembler des informations précises. Par exemple, à partir de l'adresse suivante, 'http://www.google.fr/search?hl=fr&ie=UTF-8&oe=UTF-8&q=sciences+economiques&meta=', nous relevons que l'internaute a utilisé le moteur de recherche Google dans sa version française, pour effectuer une requête intégrant les mots clés 'sciences' et/ou 'économiques'. Ce schéma descriptif correspond toutefois à une situation idéale. Parfois, certaines adresses URL sont dynamiques, c'est-à-dire que leur contenu textuel change alors que la page Internet affichée reste la même d'une visite à une autre. Cette technique est

[1] Exemple de navigateur : Microsoft Internet Explorer, Mozilla, Opera…
[2] Ou en enregistrant son contenu à des fins d'analyses ultérieures.
[3] On pourra consulter les documents RFC (*Request For Comments*) pour une description technique de l'ensemble de ces protocoles : http://www.faqs.org/rfcs.

utilisée lorsque le gestionnaire du site veut protéger le contenu d'une ou de plusieurs pages. Dans ce cas, la seule information exploitable par le chercheur concerne l'adresse racine du site.
Désormais, dans une acception plus dynamique, nous remplaçons l'adresse URL de chaque page par des lettres (exemple : 'A' pour 'http://www.education.gouv.fr' – voir la partie droite du graphique 1). Le chemin de navigation possible d'un visiteur (ou encore son parcours effectué sur le site) peut correspondre à la séquence ACEF. Autrement dit, l'internaute est arrivé sur la page A du site, puis a cliqué sur un lien pour aller sur la page C de ce même site, et de lien en lien, ce visiteur a terminé sa navigation à la page F. Remarquons que la structure du site oblige ici les internautes à suivre un chemin 'balisé'. Il faut donc avoir à l'esprit que l'étude des comportements de navigation intra-site ne peut ignorer la structure du site, c'est-à-dire la manière dont il a été construit.
Les données de navigation intra-site correspondent donc à l'enregistrement de l'adresse URL de chaque page visitée sur un même site Internet. Concrètement, les données de navigation sont des fichiers informatiques (appelés fichiers Log) contenant uniquement du texte. Ces fichiers sont la plupart du temps créés automatiquement (ils grandissent alors au fur et à mesure des visites[4]) et appartiennent au gestionnaire du site Internet[5] (appelé aussi webmaster). En dehors de certaines contraintes techniques (notamment l'existence d'un système de cache[6]), il est donc possible d'observer - à partir des fichiers Log - le chemin de navigation des visiteurs sur un site.
Nous avons jusqu'ici discuté des activités de navigation de N individus sur un site donné. Ces activités consistent à visionner une ou plusieurs pages de ce site et éventuellement répéter cette opération d'une visite à l'autre (Montgomery et al., 2004b). Puisque la navigation sur Internet consiste le plus souvent à visiter de façon séquentielle plusieurs sites, une acception plus large de l'activité de navigation considère N individus face à J différents sites. Nous observons alors les comportements de navigation inter-sites, qui peuvent là encore être enregistrés dans des fichiers texte. On parle désormais de données de navigation *inter-sites*. Ces informations sont toutefois moins communes que les données de navigation intra-site, car elles sont fondamentalement décentralisées (seule une sonde installée sur l'ordinateur de l'internaute permet d'observer ses comportements de navigation - Catledge & Pitkow, 1995).
Notons que la littérature en informatique parle souvent de 'données Log' ou de 'données serveur' lorsqu'il s'agit des données de navigation intra-site. En effet, ces dernières sont en réalité enregistrées sur des serveurs Internet. D'autre part, les données de navigation inter-sites seront davantage appelées (par cette même littérature) 'données client', car, là encore, ces dernières sont enregistrées à partir des ordinateurs personnels de chaque individu, nommés 'client'. Lorsque les

[4] Le gestionnaire du site peut décider d'effacer les données de navigation antérieures à une période donnée. Cela est souvent le cas pour les sites commerciaux, car les volumes de données alors générées présentent un coût de stockage parfois non négligeable. A l'extrême, il est possible d'effacer continuellement les données de navigation intra-site. Il suffit pour cela d'utiliser une commande informatique qui ordonne au serveur Internet d'effectuer cette opération (un serveur est un ordinateur relativement puissant qui héberge le site Internet, et gère par la suite les demandes d'informations lorsque les visiteurs se rendent sur le site).
[5] Et potentiellement tout individu qui a construit un site Internet.
[6] Nous expliquons ce phénomène très succinctement. Pour simplifier nos propos, imaginons que le réseau Internet est uniquement constitué de serveurs (équivalents à de gros ordinateurs) reliés entre eux par des fibres optiques et câbles RTC (Réseau Téléphonique Commuté). Chaque site Internet est hébergé (enregistré) sur un serveur. Dès qu'un internaute demande à visualiser la page d'un site, il effectue une requête sur le serveur hébergeant ce site, et ce dernier, en réponse, lui retourne, via le réseau Internet, la page demandée. Or, pour accélérer la vitesse de circulation de ces informations envoyées sur le réseau Internet, les pages déjà visitées sont enregistrées de façon temporaire sur l'ordinateur de l'internaute (entre autre). Ainsi, dès que cet individu demande à voir à nouveau une page, cette dernière s'affiche immédiatement sur son écran, et cela sans effectuer la moindre requête via le réseau Internet, puisqu'elle a été gardée en mémoire sur son ordinateur. Le serveur qui héberge le site et qui enregistre toutes les requêtes des internautes (pour ce site) reste donc 'aveugle' sur certains types d'actions de navigation effectuées par les visiteurs. Une restitution de la véritable séquence de navigation nécessiterait l'utilisation d'outils algorithmiques ou d'un matériel plus lourd, en l'occurrence des marqueurs (tags).

autres disciplines (notamment l'économie, le marketing ou encore la sociologie) parlent des *clickstream data*, cette appellation s'adresse la plupart du temps aux données de navigation inter-sites. Dans le cas contraire, il y a un abus de langage.
Il va de soi qu'un tel procédé d'enregistrement des comportements de navigation ne s'effectue pas à l'insu des utilisateurs. L'observation des activités de navigation inter-sites (voir même intra-site) est soumise en France à la loi 'informatique et liberté' (nommée aussi Loi sur l'Economie Numérique - LEN)[7]. Malgré cela, il faut bien être conscient que de nombreux modèles économiques sur Internet fonctionnent sur une logique d'exploitation des données de navigation. Cette dernière peut parfois être illégale en fonction de la loi en vigueur dans chaque Etat (voir à ce propos le débat relatif aux 'logiciels espions', les *spywares*).
En dehors de cela, et dans un cadre légal, la communauté scientifique dispose potentiellement d'une nouvelle source d'informations pour étudier les usages sur Internet. Voyons désormais de quelle façon ces données ont été exploitées dans la littérature et quels sont les principaux résultats relatifs aux comportements de navigation.

3.2 La littérature circonscrite à l'analyse des données de navigation et les résultats des analyses empiriques.

Les données de navigation intra-site ou inter-sites permettent de construire des variables permettant de caractériser les comportements des internautes. Ces variables sont le plus souvent le nombre de clics ou de visites sur les sites Internet. Elles ont d'abord été utilisées par les chercheurs en informatique, principalement pour tenter de résoudre les problèmes de congestion des flux d'information sur le réseau Internet. En effet, si ces chercheurs arrivaient à identifier des comportements de navigation réguliers, il devenait possible d'anticiper ces derniers et de modifier en conséquence certaines caractéristiques techniques de l'Internet (à savoir la gestion des fichiers cache, la structure des sites Internet, ou encore la conception des navigateurs), pour, au final, optimiser la vitesse de circulation des informations sur Internet. A l'extrême, plusieurs modélisations mathématiques des comportements de navigation ont été proposées (Abdulla, 1998). Depuis Glassman (1994), considéré comme précurseur dans l'analyse des données de navigation, nombre d'études dans cette lignée ont été proposées par d'autres chercheurs en informatique (par exemple, Catledge & Pitkow, 1995 ; Cunha, Bestravos , Crovella, 1995 ; Almedia et al., 1996 ; Tauscher, 1996 ; Crovella, Taqqu, Bestravos, 1998 ; Barford et al., 1998 ; Arlitt, 2000).
Si cette littérature répond principalement à des préoccupations techniques, elle peut néanmoins nous permettre d'observer certains faits saillants relatifs aux comportements des internautes.
Les principaux résultats de cette littérature empirique chargée d'analyser les comportements de navigation sont résumés dans le tableau 1 (page suivante).

[7] Voir à ce propos le site de la CNIL, http://www.cnil.fr

Tableau 1: Faits saillants sur les comportements de navigation

Résultats communs	Auteurs	Observations
Distribution identique des comportements de navigation Une majorité d'internautes visitent une minorité de sites (pages) *versus* une minorité d'internautes visitent beaucoup de sites (pages). Peu de sites (pages) reçoivent beaucoup de visites *versus* beaucoup de sites (pages) reçoivent peu de visites.	Abdulla (1998), Arlitt (2000), Adamic & Huberman (2002).	Environ 25 % des sites Internet reçoivent 80 à 90 % du total des visites en ligne (on parle souvent de la loi des 80/20).
	Glassman (1994), Catledge & Pitkow (1995), Cunha et al. (1995), Tausher (1996), Almeida et al. (1996).	Distribution fortement asymétrique du nombre de visites par page : 60 % des pages ont été visitées une fois, 19 % des pages ont été visitées deux fois, 8 % ont été visitées trois fois, 4% ont été visitées quatre fois…
Hétérogénéité des usages en ligne	Lebart & Beaudouin (2003).	Environ 14 % des internautes font 50 % des sessions. Puis 50 % des internautes effectuent 90 % des sessions.
	Montgomery & Faloutsos (2000).	Le nombre moyen de sessions par mois et par individu est de 8.1, mais la variance est très forte. Le nombre moyen de pages vues par session (et par internaute) est de 93, alors que la médiane est de 48.
Une majorité de visites répétées Forte inertie pour une majorité d'utilisateurs : la plupart des internautes visitent de façon répétée une minorité de sites Internet.	Lebart & Beaudouin, (2003).	Environ 68 % des internautes ont utilisé un seul moteur de recherche durant leur session. Concernant les sites visités, 50 % des sondés ont visité au moins une fois un site marchand dont la thématique est le tourisme ou la vente de biens culturels. Toutefois, 80 % des internautes consultent un unique site de vente de biens culturels par session, alors que près de 48 % des individus visitent plusieurs sites marchands de tourisme dans une même session.
	Glassman (1994), Catledge & Pitkow (1995), Cunha et al. (1995), Tausher (1996), Almeida et al. (1996).	Environ 58 à 61 % de l'activité de navigation consiste à visiter une page Internet déjà visualisée. La probabilité de retourner sur une même page est donc très élevée. Ce taux est indépendant de l'intensité de navigation, mais corrélé négativement à la date de la dernière visite.
Activités de recherche qui priment La principale activité de navigation concerne la recherche d'information.	Catledge & Pitkow (1995), Tausher (1996).	Les évènements de navigation les plus importants concernent la consultation de site (environ 52 % des activités de navigation), puis l'utilisation du bouton 'Back' du navigateur (environ 34 % des activités totales de navigation). Ces deux actions représentent au total environ 86% des activités de navigation. Les favoris sont rarement utilisés (< 3 %). En d'autres termes les internautes sont avant tout des chercheurs d'information.
	Lebart & Beaudouin, (2003).	93 % des internautes ont utilisé au moins une fois un moteur de recherche durant l'année.
Comportements intra-site identiques Navigation intra-site limitée pour une majorité d'individus.	Catledge & Pitkow (1995), Tausher (1996), Lukose & Huberman (2001).	A l'intérieur d'un site, la longueur moyenne du parcours de lien en lien est de 2,98 avec un écart type de 6,24.

Nous voyons alors que les études réalisées à différentes périodes et pour des échantillons de tailles non similaires, ont mis en valeur une série d'observations communes. Catledge & Pitkow (1995), Tausher (1996), Lebart & Beaudouin (2003), ont par exemple montré que la première activité de navigation sur Internet concernait la recherche d'informations.

Quant aux comportements de navigation intra-site ou inter-sites en tant que tels, la littérature a soulevé un certain nombre de régularités, alors même que l'Internet est vu comme un système complexe (Huberman, 2001) et chaotique : « *Given the dynamic nature of the web, it may be surprising for some readers to find that many properties of the Web follow regular and predictable patterns that have not changed in form over the web's lifetime* » (Pitkow, 1998, p.1).

En effet, Glassman (1994), est l'un des premiers chercheurs à avoir révélé l'existence de comportements de navigation génériques à partir des données de navigation. Sur une période d'un an, l'auteur a ainsi observé à partir d'un échantillon de 600 internautes, que la probabilité de choisir la $n^{ième}$ page la plus visitée était proportionnelle à $1/n$ (un total de 80 000 pages ont été vues par l'échantillon).

Parmi les publications ultérieures, Adamic & Huberman (2000) ont de la même manière observé des comportements identiques, en étudiant, non plus le nombre de pages visitées, mais désormais, les visites des sites Internet. Leur analyse porte sur les comportements de navigation d'environ 24 000 clients du fournisseur d'accès AOL pendant la journée du 5 décembre 1997. L'échantillon a alors visionné 3 247 054 pages sur 1 090 168 sites Internet. Les auteurs ont alors montré que 0,1 % des sites considérés (soient les 120 premiers sites les plus visités) capturaient 32,36 % des 24 000 visiteurs et que 50 % des sites rassemblaient 95 % de la totalité des visiteurs.

Sur une période d'observation plus large, et avec un plus grand nombre d'individus, Montgomery & Faloutsos (2000) ont pu observer les comportements de navigation de 74 000 individus représentatifs des ménages américains. L'enquête s'étend sur une période de 30 mois (de juillet 1997 à décembre 1999) et environ 290 millions de clics ont été enregistrés à partir des ordinateurs personnels de chaque membre d'un panel d'internautes (c'est un panel MediaMetrix - aujourd'hui devenu Nielsen/NetRatings - construit à partir de la sonde 'PC Meter'). Les usages sont mesurés en terme de temps passé à naviguer sur Internet, mais aussi en nombre de clics, de pages vues, de visites et de sessions. Là encore, une série de comportements de navigation déjà observés dans la littérature ont été relevés.

Les résultats précédents, à savoir ceux de Glassman (1994), Adamic & Huberman (2000), puis Montgomery & Faloutsos (2000) ainsi que ceux reportés dans le tableau 1, ont tous un point commun : ils peuvent être représentés par une loi de probabilité fortement asymétrique (la littérature anglo-saxonne parle de '*heavy tailed distributions*' ou encore de '*skewed distributions*' ayant une forme de J inversé), caractérisée par une moyenne supérieure à la médiane (ou encore un écart type supérieur à la moyenne du caractère étudié). Cette distribution est souvent nommée loi de Zipf (*Zipf law*) chez Glassman (1994) ou Montgomery & Faloutsos (2000), mais elle peut aussi être appelée loi puissance (*power law*) ou parfois loi de Pareto (*Pareto law*)[8]. On pourra consulter l'article de Adamic (2000) pour le lien mathématique entre ces trois distributions.

Lorsqu'une variable est distribuée selon de telles lois, cela implique que les phénomènes de petite envergure sont extrêmement communs, alors que ceux de grande envergure restent rares. Au-delà d'Internet, de nombreux exemples existent dans la 'nature', parmi lesquels la magnitude des tremblements de terre : il y a beaucoup de tremblements de terre de faible magnitude, et seulement quelques séismes de forte magnitude. En géographie, la distribution de la surface des îles d'un archipel suit systématiquement une loi puissance : il y a peu de grandes îles et beaucoup de petites. Au-delà des phénomènes naturels, Zipf (1949) montra que dans tout livre, seulement quelques mots revenaient très fréquemment (par exemple 'le' ou encore 'un'), alors que beaucoup d'autres avaient une fréquence relativement faible. Enfin, en 'économie', Pareto (1896) relevait

[8] D'autres distributions asymétriques ont été observées, par exemple la loi Gamma, ou la loi gaussienne inverse, mais les lois de Zipf, de Pareto et la loi puissance restent les plus communes lorsque l'on étudie les comportements de navigation sur Internet.

qu'un grand nombre d'individus percevaient un faible revenu, alors que l'on rencontrait très peu de fortunés.
Plus généralement, quel que soit le phénomène étudié, s'il existe une relation inverse entre la grandeur d'une variable et son occurrence (à tel point qu'une représentation Log-Log dessine un nuage de point rectiligne de pente négative proche de 1), on peut supposer l'existence d'une loi puissance pour la distribution de la variable considérée.
Puisque beaucoup de phénomènes (naturels ou non) peuvent résulter sur une distribution asymétrique, nous devinons que la littérature théorique sur un tel sujet est pléthorique. On pourra, à ce propos consulter le survey de Mitzenmacher (2003). Il existe d'autre part de nombreux modèles mathématiques (probabilistes) permettant de restituer ces lois asymétriques (Simon, 1955). En économie géographique , les modèle de Gabaix (1999) et celui de Cordoba (2001) permettent de s'écarter des considérations purement probabilistes pour intégrer une fonction d'utilité et restituer par là une distribution de Zipf relative à la taille des villes. Nul doute que l'économie de l'Internet s'est intéressée à ce type de modèle, à partir du moment où l'on observe des lois asymétriques en ligne (Adamic & Huberman, 2000). De ce point de vue, la recherche qui s'intéresse à la modélisation (économique) des comportements de navigation n'en n'est qu'à ses débuts. Notre approche reste néanmoins ici empirique (la littérature est dans ce cas plus volumineuse), mais elle ne remet pas en cause une réflexion théorique relative aux modélisations possibles des comportements de navigation[9].
La prochaine section propose d'utiliser une base de données de navigation pour construire des lois asymétriques.

4. Une analyse descriptive des comportements de navigation

4.1 L'échantillon : la base de données Boston University Web Client

L'expérience 'BU Web Client' consistait à enregistrer les données de navigation d'un échantillon d'individus sur plusieurs mois. Le département informatique de l'université de Boston s'est chargé de cette opération qui fait partie d'un projet de recherche plus vaste appelé OCEAN (Objet Caching Environments for Applications and Network Services)[10]. Ce projet consistait à trouver un mode optimal de gestion des fichiers cache sur Internet, problème lié à l'informatique et non pas aux Sciences Economiques. Dans ce sens, aucune technique d'échantillonnage n'a eu lieu. Notons aussi que les informations utilisées datent de 1995. Nous sommes donc aux prémices de l'Internet. Ces deux insuffisances qui paraissent pesantes, doivent toutefois être minorées :

- tout d'abord, le mérite des initiateurs de l'expérience 'BU Web Client' est d'offrir gracieusement leurs données de navigation à la communauté internationale. Ces données sont aujourd'hui les seules disponibles (à ce niveau de qualité). Elles ont en effet été enregistrées à partir des ordinateurs utilisés par chaque membre de l'échantillon (en modifiant les caractéristiques du navigateur MOSAIC, logiciel libre et dominant sur le marché en 1995). Ce sont donc des données dites 'clients' qui permettent aussi bien d'observer les comportements de navigation intra-site que les comportements inter-sites. A notre connaissance, ces informations n'ont jamais été utilisées à des fins d'analyse en Sciences Economiques,

[9] Dans une acception 'économique', nous supposons que la navigation consiste à visiter une série de sites Internet. Chaque site représente un choix possible pour l'individu. L'internaute visite alors le site qui maximise son espérance d'utilité. Nous supposons d'autre part que l'internaute possède un capital limité de visites sur chacun des sites. Cela se justifie par le fait qu'une ressource rare sur Internet est l'attention (le temps) de l'internaute. Nous sommes donc dans un cadre de maximisation d'une fonction d'utilité (intertemporelle) sous contrainte de temps (ou encore d'attention), (Le Guel, 2004).
[10] http://cs-www.bu.edu/groups/oceans/Home.html

• ensuite, si l'échantillon n'est pas représentatif du point de vue du profil des internautes (nous n'avons d'ailleurs aucune information à ce propos), il peut l'être du point de vue du profil des comportements de navigation. En effet, il n'y a aucune raison de penser que les 'lois de surfing' observées dans la littérature sur des échantillons différents et à des périodes non semblables, ne s'appliquent pas à notre échantillon. Nous supposons donc que les comportements de navigation de nos individus doivent être distribués selon une loi asymétrique telle que la loi de Zipf, la loi de Pareto ou la loi puissance. Si cela est vérifié, nous admettons que notre échantillon est viable, dans la mesure où nous confirmons les résultats déjà observés dans les études empiriques préalables. Nous pourrons alors dépasser cette littérature essentiellement issue de l'informatique, pour nous attacher à des considérations davantage économiques.

Il y a à l'origine 762 étudiants, formant deux échantillons distincts. Le premier échantillon est composé de 214 étudiants niveau licence/maîtrise. Le second échantillon contient 548 élèves niveau première année/DEUG. Nous avons éliminé l'échantillon des étudiants de licence/maîtrise (équivalent). Les données montraient en effet de nombreuses faiblesses dues à des problèmes techniques (indépendants de notre volonté). Notre échantillon initial est donc composé de 548 étudiants volontaires de l'université de Boston. Nous avons sélectionné uniquement les élèves ayant navigué en dehors du site de leur université (il était en effet possible d'utiliser exclusivement le réseau intranet), soit 467 individus. Nous avons d'autre part choisi les enregistrements des mois de janvier et février 1995, les autres mois présentaient des incohérences (les informaticiens ont en effet réalisé plusieurs tests sur le réseau de l'université de Boston, ce qui bruite les observations, excepté pour les mois de janvier et février).
Les caractéristiques des comportements de navigation de l'échantillon des 467 étudiants sont présentées dans le tableau 2 (ci-dessous).

Tableau 2 : Caractéristiques de navigation des étudiants (467 individus)

	Nombre de sessions par individu	**Nombre de visites par individu**
Minimum	1	1
Maximum	174	327
Total (pour les 467 individus)	5348	33100
Espérance	11.45	70.18
Ecart type	18.32	143.30

Pour simplifier le tableau et sans perte de généralité, nous présentons uniquement le nombre de sessions et de visites (il était par exemple possible de travailler sur le volume de pages vues). Ces informations sont obtenues après un traitement séquentiel et relativement lourd des informations (une centaine de programmes informatiques écrits et plusieurs jours de 'calculs' sur ordinateur) .
Le traitement s'effectue en deux étapes majeures. La première étape englobe les opérations de reconnaissance de mots pour chaque ligne des fichiers texte. Nous avons compté pour notre échantillon environ 200 000 lignes. La seconde étape consiste à dénombrer l'ensemble de ces mots afin de construire nos statistiques.
Une session est définie comme la période entre laquelle un individu se connecte puis se déconnecte pour naviguer sur Internet. Durant chacune des sessions, un internaute peut aller visiter un ou plusieurs sites. Dès lors, le nombre de sessions est inférieur ou égal au nombre de visites. Notons d'autre part que les sessions et les visites peuvent se situer à des périodes différentes pour chacun des étudiants, certains se connectant en début de semaine ou aux heures des repas, d'autres préférant naviguer sur Internet le week-end ou le soir. La seule contrainte reste l'intervalle de temps de deux mois (janvier et février 1995), commune aux membres du panel. Nous remarquons immédiatement des écarts-types supérieurs aux espérances. Cela est souvent signe de surdispersion et par conséquent d'hétérogénéité entre les individus, en termes de nombre de visites sur une même période ou de nombre de sessions.

4.2 La construction d'une loi de Zipf

Une fois cette première 'photographie' des comportements en ligne réalisée, nous devons aller plus loin dans la description des profils de navigation. Il n'existe pas encore de procédures strictes d'analyses descriptives pour ce type de données. De nombreux résultats peuvent toutefois être proposés, notamment l'identification d'une loi de Zipf relative au nombre de visites sur chaque site Internet.

Nous supposons qu'une visite débute lorsqu'un internaute se présente sur un site pour visionner une ou plusieurs pages et s'achève lorsque les deux conditions suivantes sont vérifiées en même temps :

1. lorsque l'individu poursuit sa navigation sur un autre site ou décide d'arrêter sa connexion,
2. et lorsque l'internaute n'a pas navigué sur le site considéré depuis plus de 30 minutes.

La seconde condition permet de diminuer le nombre de visites répétées sur un même site dues à des 'allées et retours' systématiques et très courts (en terme de temps) de ce site vers d'autres sites. Cela nous permet d'améliorer la mesure du niveau de fidélité (de persistance) ou d'intérêt de l'internaute vis-à-vis d'un site donné.

A priori, il n'y a pas lieu de penser que la distribution du nombre de visites sur chaque site ne suive pas une loi normale. Dit autrement, le nombre moyen de visites par site devrait correspondre à l'occurrence la plus forte. Malgré cela, nous avons vu précédemment, que les distributions statistiques les plus fréquentes sur Internet étaient fortement asymétriques (Adamic & Huberman, 2000, 2002 ; Montgomery & Faloutsos, 2000). Dans ce schéma, l'évènement moyen n'a pas la probabilité d'apparition la plus élevée. En d'autres termes, les évènements de grande ampleur sont peu probables, alors que les évènements de petite ampleur le sont plus.

Afin d'observer les comportements de navigation de notre échantillon et valider l'hypothèse selon laquelle une distribution asymétrique a lieu (via l'observation d'une loi de Zipf, d'une loi de Pareto ou encore d'une loi puissance, ces trois distributions étant, rappelons-le, voisines – Adamic 2000), nous allons estimer les paramètres de la première loi, à savoir la distribution de Zipf.

La loi de Zipf (1949), fait correspondre la taille d'un évènement, ici le nombre de visites sur chaque site (noté V), au rang de cet évènement, assimilé dans ce chapitre au classement des sites Internet en nombre de visites (noté R), tel que $V = R^{-z}$ avec z la pente proche de 1. En d'autres termes, selon la loi de Zipf, la taille du $R^{ème}$ site Internet devrait être inversement proportionnelle à son rang (notons que V peut par la suite être exprimé en terme de fréquences).

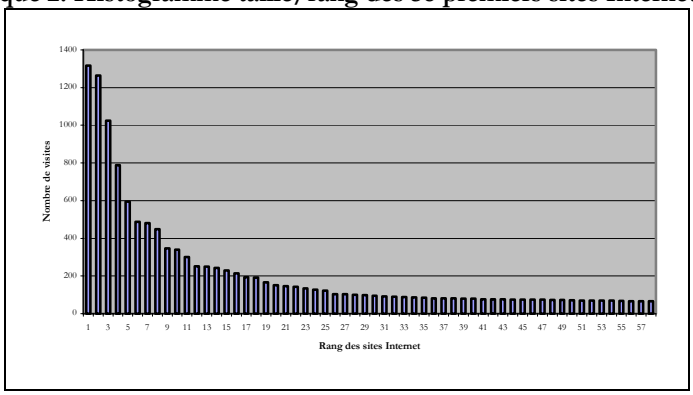
Dans notre échantillon, le nombre total de sites Internet visités une fois ou plus par nos 467 étudiants pendant les deux mois de navigation est de 5772. Le nombre total de visites est de 33100. En moyenne, chaque site a donc été visité 5,73 fois, mais la distribution est en réalité très inégale puisque l'écart type du nombre de visites est environ 6 fois supérieur à la moyenne. D'autres statistiques descriptives sont fournies dans le tableau 3.

Tableau 3: Statistiques descriptives pour la distribution des visites de sites Internet

Nombre de sites Internet visités	5772
Nombre de visites	33100
Nombre minimum de visites sur un site	1
Nombre maximum de visites sur un site	1317 (moteur de recherche Yahoo)
Médiane	1
$3^{ème}$ quartile des visites	3
Moyenne des visites pour chaque site [IC à 95 %]	5,735 [4,824 ; 6,645]
Ecart type	35,283
Coefficient d'asymétrie de la distribution	25,679

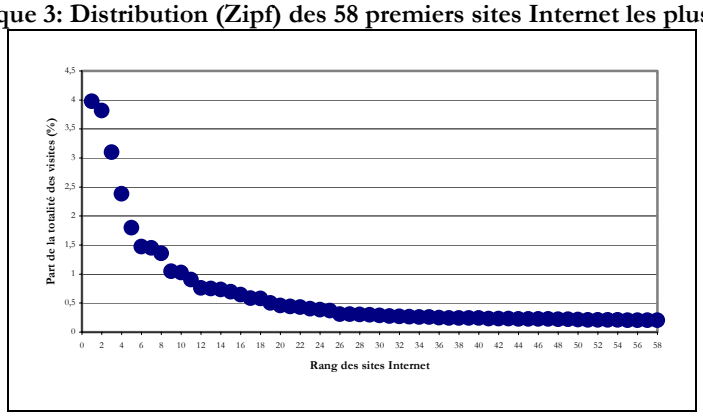
Le graphique 2 propose la distribution taille[11]/rang des 58 premiers sites les plus visités parmi les 5772. L'analyse graphique confirme les résultats issus des statistiques descriptives : la distribution des visites pour chaque site Internet est très asymétrique.

Graphique 2: Histogramme taille/rang des 58 premiers sites Internet visités



Le graphique montre que le premier site (qui est en réalité le moteur de recherche éducatif Yahoo[12]) rassemble 1317 visites, alors que le 58ème ne compte plus que 67 visites. Nous avons calculé que l'ensemble de ces 58 premiers sites Internet (soit environ 1% de la totalité des 5772 sites visités) représente pratiquement 40 % des 33100 visites effectuées sur la période d'observation.
Souvent, la distribution de Zipf est présentée comme suit dans la littérature (graphique 3), l'axe des ordonnées laissant parfois place aux occurrences, les 'bâtons' du graphique étant d'autre part transformés en points.

Graphique 3: Distribution (Zipf) des 58 premiers sites Internet les plus visités



Nous voyons par exemple sur ce graphique que le site Internet le plus visité représente environ 4 % des 33100 visites effectuées sur 5772 sites différents. Le tableau 4 détaille l'adresse des 10 premiers sites les plus visités par notre échantillon.

Tableau 4 : Occurrences des visites pour les 10 premiers sites

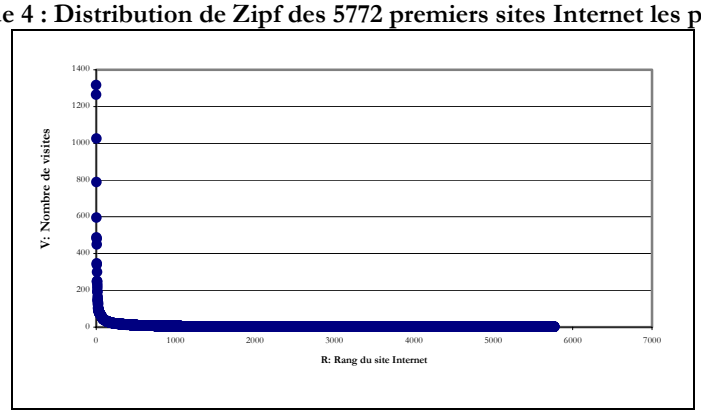
Rang	Part des visites totales	Adresse du site	Thème du site
1	3,98 %	http://akebono.stanford.edu/	Moteur de recherche Yahoo éducatif
2	3,82 %	http://www.yahoo.com/	Moteur de recherche Yahoo
3	3,10 %	http://www.ncsa.uiuc.edu/	Education
4	2,38 %	http://home.mcom.com/	Moteur de recherche
5	1,80 %	http://nearnet.gnn.com/	Moteur de recherche

[11] En nombre de visites par site.
[12] Qui, à ce jour, n'existe plus sous sa forme de 1995.

6	1,47 %	http://info.cern.ch/	Centre de recherche
7	1,45 %	http://sunsite.unc.edu/	Education
8	1,36 %	http://www.mit.edu:8001/	Education
9	1,05 %	http://www.w3.org/	Consortium Internet
10	1,03 %	http://www.timeinc.com/	Site de la Time Warner

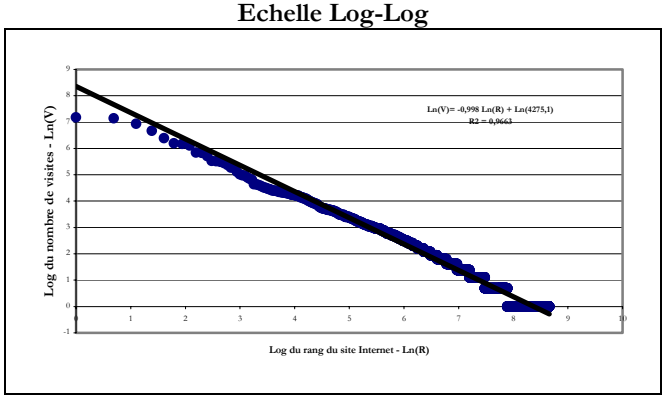
Les moteurs de recherche ainsi que les sites éducatifs représentent une part majeure des visites. Le graphique 4 élargit désormais la distribution taille/rang des visites à la totalité des 5772 sites Internet visionnés une fois ou plus par l'échantillon. La distribution est si extrême qu'elle dessine une parfaite hyperbole rectangulaire.

Graphique 4 : Distribution de Zipf des 5772 premiers sites Internet les plus visités



Pour vérifier l'existence d'une loi de Zipf, la méthode la plus utilisée consiste à effectuer un ajustement linéaire (par les moindres carrés ordinaires) sur la distribution taille/rang, en exprimant chacun des axes sous leur forme logarithmique[13]. L'expression mathématique de la loi de Zipf devient $Ln(V) = -z \quad Ln(R) + Ln(c)$ avec c une constante additionnelle qui ne modifie en rien la forme de la distribution. Le graphique 5 présente la loi de Zipf sur une échelle Log-Log. La valeur de la pente estimée est de -0,998. Nous avons donc un rapport inverse presque parfait entre le nombre de visites (la taille) et le rang du site. La distribution semble être proche d'une loi de Zipf de pente -1.

Graphique 5 : Distribution de Zipf des 5772 premiers sites Internet les plus visités
Echelle Log-Log



[13] Une méthode alternative permet de tester statistiquement l'existence d'une loi de Zipf (Urzùa, 2000). Toutefois une controverse importante demeure dans la littérature. Selon Gabaix (1999), lorsque la taille de l'échantillon augmente, la pente de la loi de Zipf tend vers -1. A contrario, selon le test de Urzùa, lorsque la taille de l'échantillon augmente, nous nous éloignons de cette pente théorique. Face à cette controverse, nous choisissons de suivre la technique d'estimation la plus utilisée, à savoir l'ajustement linéaire.

4.3 La construction de la loi de Pareto et de la loi puissance

A partir des mêmes données de navigation, il est possible de construire d'autres distributions asymétriques (Pitkow, 1998). Parmi ces dernières, la loi puissance et la loi de Pareto (1896) demeurent les plus utilisées dans l'analyse des comportements de navigation. L'encadré détaille la formulation mathématique de ces deux lois.

Une variable aléatoire X suit une loi de Pareto lorsqu'elle satisfait l'égalité suivante :

$$\Pr[X \geq x] = \left(\frac{x}{m}\right)^{-p},$$

où $m > 0$ est la valeur minimale de x (cette loi a d'abord été utilisée pour observer la distribution des revenus dans la population, dès lors m correspondait au revenu minimum), avec $x \geq m$ et $p > 0$, la pente de la loi de Pareto. Nous en déduisons la distribution cumulée :

$$\Pr[X < x] = 1 - \Pr[X \geq x]$$

Dès lors, la dérivée première en x de la loi cumulée conduit à la densité de probabilité suivante :

$$\Pr[X = x] = pm^{p}x^{-(p+1)},$$

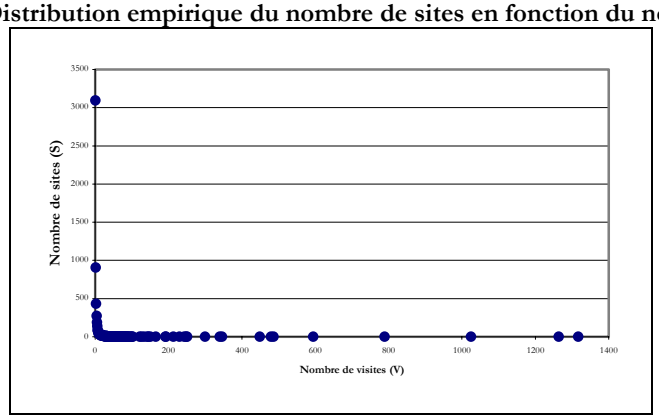
La formule précédente correspond à la distribution généralisée de Pareto qui est aussi appelée loi puissance. En d'autres termes, la loi puissance n'est rien d'autre que la fonction de densité de la distribution de Pareto.

La loi de Pareto et la loi puissance étant en réalité identiques, l'estimation d'une d'entres elle suffit. Appliquée à nos données, la loi puissance met en relation le nombre de sites Internet S au nombre de visites V (graphique 6) tel que :

$$S = aV^{-w} ;$$

avec, $a > 0$ une constante, $w > 0$ la pente de la loi puissance.
Notons que $w \approx p + 1$. Lorsque w est proche de 2, on admet que la variable aléatoire V suit une loi puissance (Simon, 1955 ; Adamic, 2000 ; Mitzenmacher, 2003). D'autre part, S peut être exprimée en termes de fréquences ou d'effectifs, seule la valeur de a est modifiée, mais en aucun cas celle de w.
Pour construire notre distribution, nous devons compter le nombre de sites Internet qui ont reçu une seule visite, ceux qui ont eu deux visites, et ainsi de suite. Là encore, nous voyons sur le graphique 6 que la relation nombre de sites/nombre de visites est très asymétrique. Par exemple, 3093 sites ont connu une seule visite sur la période d'observation (la fréquence empirique correspond à 54 % environ). Ensuite, 907 sites ont eu deux visites (environ 16 % de la totalité des sites). A l'extrême, un seul site (0,002 % de l'échantillon) a connu 1317 visites.

Graphique 6 : Distribution empirique du nombre de sites en fonction du nombre de visites



Par la suite, l'estimation de la pente w peut s'effectuer comme pour la distribution de Zipf ou de Pareto, c'est-à-dire en exprimant la loi puissance sous sa forme logarithmique, afin d'implémenter un ajustement linéaire par les moindres carrés ordinaires :

$$\ln(S) = - w \ln(V) + \ln(a)$$

En suivant la méthode d'estimation de Adamic (2000) (utilisation d'intervalles à amplitude exponentielle pour le nombre de visites) et en respectant la règle de proportionnalité de la surface des bâtons de l'histogramme, nous obtenons une pente w pour la loi puissance de -2,07 (pour plus de détails ainsi que l'estimation de la pente de la distribution de Pareto, voir Le Guel, 2004).
In fine, nous vérifions la correspondance des pentes des distributions de Zipf (z), de Pareto (p) et de la loi puissance (w) tel que :

$$\frac{1}{z} + 1 \text{ » } w \text{ » } 1 + p \quad \text{(Adamic, 2000).}$$

Relativement à nos estimations :

$$\frac{1}{z} + 1 \text{ » } w \text{ » } 1 + p \text{ Þ } \frac{1}{0,998} + 1 \text{ » } 2,07 \text{ » } 1 + 1,109 \text{ Þ } 2,002 \text{ » } 2,07 \text{ » } 2,109$$

Le calcul des pentes nous permet donc de vérifier un résultat désormais classique de la littérature. Notre échantillon répond à des comportements de navigation 'dits invariants', dans le sens où ils peuvent être décrits sous la forme d'une distribution asymétrique.
L'estimation des différentes pentes pour notre échantillon nous permet de vérifier une 'représentativité' de notre échantillon relativement aux résultats de la littérature empirique. Notons d'autre part que l'utilisation d'une loi plutôt qu'une autre importe peu, seul le type d'information présenté dans les axes des graphiques change (prise en compte où non du 'rang' du site). Une remarque importante doit néanmoins être faite. Les résultats précédents s'intéressent uniquement aux comportements de navigation que l'on pourrait qualifier « d'agrégés ». Dans ce cas de figure, on se concentre avant tout sur les sites Internet en rassemblant les occurrences de visites des sites pour un échantillon d'internautes. Or, cela conduit selon nous à une perte d'informations importante (Bucklin & Sismeiro, 2003), car nous ne mesurons pas réellement les comportements individuels de navigation. A l'extrême, les lois puissances au niveau agrégé ne montrent qu'une fracture numérique du côté de l'offre (c'est-à-dire les sites Internet), où peu de sites reçoivent la majorité des visites. Nous avons là un 'état' de la concurrence sur Internet, mais en aucun cas une mesure potentielle de la fracture numérique des usages en ligne au sens d'Hargittai. Rappelons en effet que notre souci est d'étudier la potentialité des données de navigation pour observer une fracture numérique des usages définie comme « la capacité des individus à utiliser Internet de façon effective et efficiente ». Dans ce sens, il nous semble que les variables 'nombre de visites' ou 'nombre de sites' *par individu* peuvent être des indicateurs viables pour évaluer cette capacité à utiliser Internet. Ainsi, nous proposons dans la section suivante une analyse des comportements de navigation que l'on qualifie cette fois de *désagrégée*.

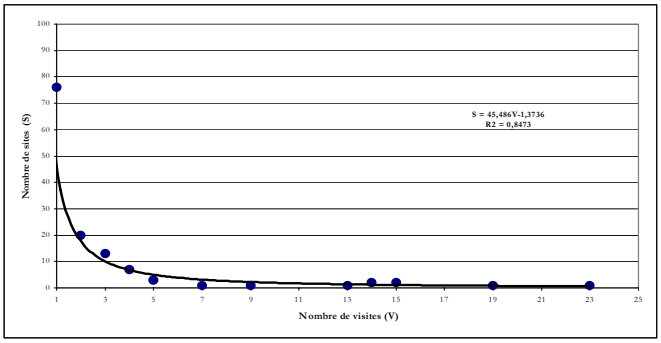
5. Proposition pour une analyse des comportements de navigation au niveau désagrégé : la construction 'des pentes individuelles' et leurs conséquences du point de vue de la fracture numérique de second niveau.

5.1 Les pentes individuelles

Seulement quelques travaux placés au niveau des individus ont permis de segmenter les internautes en fonction de leurs comportements de navigation *individuel*. Cette littérature est souvent issue de l'informatique (Catledge & Pitkow, 1995 ; Cunha, Bestavros, Crovella, 1995 ; Pitkow, 1998). Il existe néanmoins des travaux en gestion (Christ et al., 2001 ; Park & Fader, 2004). L'ensemble de ces études a résulté sur la définition d'un 'spectre comportemental' allant du 'faible' utilisateur (moins de 10 sites Internet visités par semaine, soit 35 % de l'échantillon chez Christ et al., 2001) à l'utilisateur 'acharné' (en moyenne 50 sites visités par semaine, soit 4,3 % de l'échantillon (ibid., 2001)[14].

Dans cette section, nous proposons d'utiliser les distributions asymétriques pour caractériser les comportements de navigation de chaque individu de notre échantillon. Nous choisissons alors d'appliquer la loi puissance[15] non plus au niveau agrégé, mais désormais au niveau individuel. Il est en effet possible de compter, pour un individu, le nombre de visites sur chaque site à la fin d'une période d'observation considérée. A notre connaissance, une telle démarche n'a pas été proposée dans la littérature en Sciences Economiques. Le graphique 7 présente une distribution puissance pour l'individu 'numéro 160' de notre échantillon.

Graphique 7 : Distribution puissance pour un internaute



Si nous regardons les deux points extrêmes de cette figure, nous voyons pour le quart nord-ouest, que l'internaute a visité 76 sites *différents* une seul fois, [point de coordonnée (1, 76)], avec V , le nombre de visites et S , le nombre de sites. A l'extrémité du nuage de points (quart sud-est), l'individu a visité 23 fois un même site (en réalité l'adresse http://www.timeinc.com), point de coordonnée (23, 1). Au vu de ce graphique, nous supposons alors qu'un nombre important de points dans le quart sud-est, relativement au quart nord-ouest et sud-ouest, peut être assimilé à un comportement de navigation relativement 'routinier' pour l'individu 160. En d'autres termes, ce dernier a tendance à revenir régulièrement sur peu de sites, c'est-à-dire à utiliser une faible part de l'offre de sites Internet. A l'inverse, une majorité de points dans la partie ouest du graphique stipule un comportement moins routinier, l'internaute aurait alors tendance à visiter plus de sites (certes, moins souvent), donc à utiliser davantage Internet. Or, la répartition de l'ensemble des points du graphique influe globalement sur le profil de la loi puissance qui lui ait ajustée.

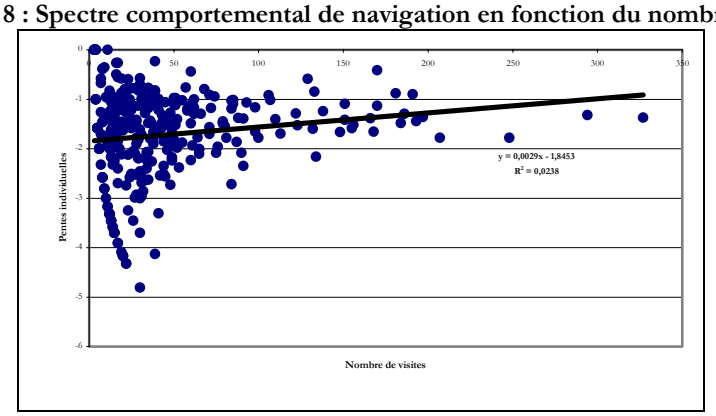
En effet, tout autant qu'au niveau agrégé, il est possible, au niveau individuel, d'estimer la pente de la loi puissance du graphique précédent. Ainsi, pour l'individu 'numéro 160', cette pente individuelle est estimée à -1,3736. Plus généralement, nous pouvons penser que le niveau de la pente de chaque internaute est corrélé à un type de comportement de navigation. La pente d'une loi puissance variant de zéro à l'infini, nous pouvons supposer que plus cette pente est proche de zéro et plus l'individu a tendance à visiter régulièrement peu de sites. Au contraire, lorsque cette pente est élevée (-5 par exemple), l'individu a tendance à visiter peu de fois un nombre important de sites Internet. Dans ce sens, nous nous rapprochons des observations de Cunha, Bestavros,

[14] Dans l'article de Christ et al. (2001), les non-utilisateurs représentent 49,9% de l'échantillon, et les utilisateurs confirmés (*heavy users*) ont une proportion de 10,2 %. L'échantillon observé est celui du projet HomeNet (http://homenet.hcii.cs.cmu.edu/progress) et la période d'observation est d'environ quatre ans (1995-1998).

[15] Rappelons qu'une loi de Zipf (ou de Pareto) aurait pu être choisie, sans remise en cause des résultats.

Crovella, 1995. Rappelons que les auteurs ont construit un indicateur permettant de restituer la capacité des internautes *à découvrir de nouveaux sites Internet*. De notre point de vue, un internaute 'routinier' aura tendance à découvrir peu de nouveaux sites. Sa capacité à utiliser Internet resterait donc limitée. La construction des lois puissances individuelles pourrait donc être une première mesure (certes très imparfaite) de la « capacité des individus à utiliser Internet de façon effective et efficiente ». Le graphique 8 reporte les pentes individuelles des lois puissance des 323 individus[16] de notre échantillon.

Graphique 8 : Spectre comportemental de navigation en fonction du nombre de visites



La pente moyenne est de -1,71 (avec un écart type égal à 0,87) et le coefficient de détermination moyen $\overline{R2}$ des ajustements linéaires pour chaque individu est de 0,86 avec un écart type de 0,16. Nous remarquons immédiatement une forte hétérogénéité des pentes individuelles, phénomène que l'on ne pouvait pas distinguer au niveau agrégé, puisque seule une pente proche de -2 était estimée pour l'ensemble de l'échantillon. Une approche désagrégée des comportements peut donc être instructive. Néanmoins, selon Gabaix (1999) la pente d'une loi puissance tend vers 2 lorsque le nombre d'observations tend vers l'infini. Pour contrôler ce phénomène, nous ajoutons en abscisse le nombre de visites. Nous voyons immédiatement que la variabilité des pentes a tendance à diminuer lorsque le nombre de visites augmente. Cela soulève la question de l'existence d'un comportement 'générique' de navigation pour les internautes, au même titre que les résultats d'Adamic et Huberman (2000). Rappelons que les auteurs ont observé, au niveau agrégé, une loi dite 'universelle' de navigation[17] (nous pesons que cette expression est galvaudée puisqu'elle ne se situe pas réellement au niveau individuel) où peu de sites rassemblent la majorité des visites. Au niveau individuel, est-il alors possible que l'ensemble des internautes ait une distribution des visites de sites identique ? La section suivante discute des conséquences d'une telle question du point de vue de l'analyse de la fracture des usages en ligne.

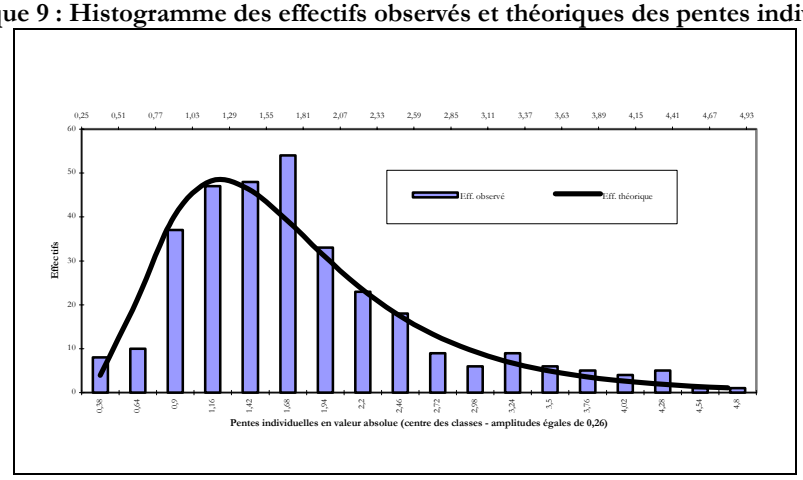
5.2 Conséquences de l'existence des lois asymétriques sur la fracture numérique des usages

Le graphique 9 (que l'on peut appeler 'spectre des comportements de navigation') propose de tracer la distribution des pentes individuelles de notre échantillon. Nous reprenons dans ce sens le graphique 8 sans prendre en compte l'axe des abscisses relatif au nombre de visites. Puisque notre échantillon est inférieur à 800 observations, nous choisissons un nombre de classes égal à $\sqrt{323}$, soit environ 18 intervalles[18]. Un premier test de normalité de la distribution[19] au seuil de

[16] Si un individu a navigué une fois sur chaque site, ou visité *n* fois un même site durant la période de l'étude, la pente ne peut être calculée. Nous n'avons donc pas pris en compte ces internautes. Cela avait peu d'incidence sur nos résultats.
[17] En anglais : la « *surfing law* ».
[18] Selon le *Stata Graphics Manual*, Release 7, p. 70.
[19] Nous avons réalisé deux tests au seuil de 1%. L'hypothèse H0 stipule que l'échantillon suit une loi normale. Le premier test est celui de Shapiro-Wilk. La valeur observée *W* est de 0,932, la p-value est inférieure à 0,0001,

1% est rejeté, l'échantillon ne suit pas une loi normale. Nous proposons alors d'ajuster une loi log-normale[20]. Un test bilatéral de Kolmogorov-Smirnov au seuil de 1% ne rejette pas l'hypothèse nulle H0 : il n'y a pas de différence entre la distribution empirique et théorique[21]. En d'autres termes, la distribution des pentes individuelles peut être correctement approximée par une loi log-normale. Un test du Khi-deux de conformité entre les effectifs théoriques et les effectifs observés confirme nos observations[22].

Graphique 9 : Histogramme des effectifs observés et théoriques des pentes individuelles



La distribution précédente nous montre que les comportements sont hétérogènes : nous ne sommes pas face à une loi uniforme, en d'autres termes, les pentes individuelles de chaque internaute ne sont pas identiques. A l'image de Catledge & Pitkow (1995) ; Cunha, Bestavros, Crovella, (1995) ; Pitkow (1998) ; Christ et al., (2001) ; Park & Fader (2004), nous pouvons définir des groupes d'internautes. De notre côté, trois groupes se dessinent. Une majorité d'individus (environ 68 % d'entre eux) ont une pente comprise entre -1 et -2. Ensuite, environ 27 % des internautes ont une pente individuelle supérieure à -2. Enfin, une minorité d'individus (6 % de l'échantillon) ont une pente inférieure à -1.

Comme nous l'avons vu, il n'est malgré tout pas certain que la distribution des pentes individuelles puisse continuer à suivre une loi log-normale, à partir du moment où l'on augmente la période d'observation des navigations (voir même la taille de l'échantillon). D'autres lois de probabilité sont en effet possibles. Si la distribution log-normale tend vers une loi puissance, la plupart des individus ont une pente faible et demeurent donc relativement routiniers. Une loi Beta stipulerait le phénomène inverse : les internautes sont majoritairement non routiniers[23]. Une distribution normale des pentes individuelles résulterait sur l'appréciation d'un comportement moyen en ligne avec un écart toléré. Enfin, si la distribution tend vers une loi uniforme, il y a convergence des pentes individuelles vers une valeur unique. Or, la valeur empirique la plus fréquente de la pente d'une loi puissance est -2. Une telle valeur pour chaque pente indiquerait que les distributions individuelles du nombre de visites sur chaque site ont un profil similaire. Dans ce cas - au même titre que Adamic & Huberman (2000), mais au niveau individuel - il existerait une loi de navigation 'universelle' où tous les internautes visiteraient régulièrement une

---

H0 est donc rejetée. Le second test est celui de Jarque-Bera à deux degrés de liberté. Le JB observé est de 63,3, le JB critique est 13,8. La p-value est inférieure à 0,0001, H0 est rejetée.

[20] D'espérance 0,445 et de variance 0,247.

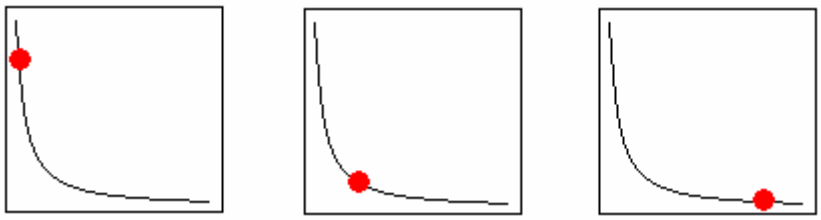
[21] La valeur observée de Kolmogorov-Smirnov est de 0,048, la valeur critique est de 0,858 et la p-value est de 0,453.

[22] Au seuil de 1%, le Khi-deux observé est de 27,206, le Khi-deux critique est de 37,697 et la p-value est de 0,027.

[23] Pour a >1 et b=0,5.

minorité de sites Internet. Pour vérifier une telle hypothèse, il faudrait, comme nous l'avons vu, augmenter la durée d'observation de l'échantillon d'internautes.
Si désormais, nous acceptons l'hypothèse selon laquelle tous les internautes distribuent leurs visites en ligne selon une loi puissance de pente -2, est-il possible qu'une telle observation puisse invalider l'existence d'une fracture numérique des usages en ligne, puisque tout le monde se comporterait de la même manière ?
Selon nous, rien n'est moins sûr. Il y a en effet plusieurs raisons à cela.
Tout d'abord, même dans ce cas de figure (si les pentes individuelles sont toutes identiques), une hétérogénéité des comportements de navigation demeure. Pour s'en rendre compte, il suffit d'observer le graphique stylisé suivant (graphique 10) et d'imaginer dans un premier temps que les lois puissances individuelles (de pente identique) s'adressent à trois internautes. Il y a donc une distribution pour chaque individu. Si le profil de chaque distribution est identique, rien ne dit que le 'portefeuille' de sites Internet visités par chacun des trois individus soit le même. Derrière cette homogénéité des distributions individuelles, il est donc possible d'observer une forte hétérogénéité dans la 'consommation' des sites en ligne.

Graphique 10 : lois puissances individuelles stylisées



Ensuite, même si nos trois internautes ont utilisé un même site (représenté sur le graphique 10 par un point apposé le long des lois puissances respectives), chacun d'entre-eux ne le consomme pas de la même manière, l'individu 3 (à droite) étant plus routinier que l'individu 2 (au milieu) qui est lui-même plus routinier que l'individu 1 (à gauche). Cette hétérogénéité peut être source d'inégalité (de fracture) lorsque le site considéré est utile socialement ou lorsqu'il permet à un individu d'en tirer une information importante (par exemple afin d'acheter un bien sur Internet au plus bas prix).
Enfin, il faut comprendre qu'une distribution asymétrique telle que la loi puissance est le résultat, *à un moment donné*, d'un processus dynamique complexe où l'internaute se trouve constamment face à deux choix principaux lors de sa navigation : visiter un nouveau site Internet ou revenir sur un site qu'il connaît déjà. Dès lors, l'observation - *à un moment donné* - d'une loi puissance pour un individu ne donne aucune information sur la dynamique de ses choix de visites en ligne. Or, c'est peut être dans cette dynamique des choix individuels que l'on peut trouver les véritables sources d'une fracture des usages en ligne. En effet, il est selon nous important de pouvoir distinguer les raisons d'un comportement routinier. Faut-il voir une incapacité de l'internaute à partir d'un site Internet (pour aller voir la concurrence par exemple) où à découvrir de nouveaux sites, où faut-il conclure que l'internaute est pleinement satisfait de l'ensemble de ses choix de sites ? Dans le second cas, une fracture des usages en ligne au sens d'Hargittai est moins prégnante.
Finalement, l'utilisation des données de navigation devrait permettre de mieux discerner la part d'hétérogénéité des comportements due à une incapacité à utiliser Internet de façon effective et efficiente. Une telle démarche est encore aujourd'hui à l'état de projet pour les disciplines qui s'intéressent à ce type de problématique.

6. Conclusion :

Cette contribution a voulu répondre à deux questions. La première concerne la mesure des usages en ligne. Nous avons proposé d'utiliser les données de navigation pour décrire les comportements de visites des internautes sur chacun des sites. Il semble que ce nouveau type de données puisse aider à tester l'hypothèse d'Hargittai (2002) selon laquelle il existe une fracture numérique de second niveau ancrée dans la capacité des individus à utiliser Internet.
La seconde question concerne l'existence d'une loi de navigation 'universelle' au niveau des internautes. De ce point de vue, l'utilisation des données de navigation nous a permis d'observer une forte hétérogénéité des comportements de visites sur chaque site, d'un individu à un autre. Cette dernière reste toutefois à vérifier. Il se peut que sur une période d'observation plus longue, les comportements de visites de chaque internaute se distribuent selon une loi asymétrique. Néanmoins, même dans ce cas de figure, une hétérogénéité demeure puisqu'il est peu probable que tous les internautes visitent un portefeuille identique de sites. En admettant toutefois qu'un tel phénomène se créer, là encore, une hétérogénéité des comportements demeure, car la dynamique des visites pour chaque internaute (capacité à visiter/découvrir un site et à en partir) doit être variable. C'est alors dans l'étude de cette dynamique que l'on pourrait supposer que l'hétérogénéité des comportements exprime une inégalité dans les usages.
Finalement, la question de la dynamique des visites (des choix) sur Internet nous fait penser à la problématique de l'adoption de la diffusion d'un bien innovant. La diffusion d'une innovation ou d'un service rend compte de la distribution temporelle des décisions d'adoption d'une nouvelle technologie au sein d'une population donnée d'individus (système social) ou d'entreprises : « *Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system* » (Rogers, 1995, p. 5). Au niveau agrégé (pour l'ensemble de la population), le schéma empirique d'adoption d'une innovation (produit ou service) à travers le temps correspond souvent à une courbe sigmoïdale (en forme de '*S*'). De nombreux modèles mathématiques ont été implémentés pour prévoir la diffusion d'une innovation à partir d'une somme d'informations minimales relatives aux premières périodes de l'adoption (Mahajan, Peterson, 1985 ; Curien & Gensollen, 1989 ; De Palma, Droesbeke, Lefèvre, 1991 ; Fildes & Kumar, 2002). Fondamentalement, ces modèles supposent que chaque adoptant potentiel est influencé par des facteurs dits 'internes' (c'est-à-dire via les phénomènes de bouche à oreille ou d'imitation) (Mansfield, 1961), et/ou des facteurs 'externes' (par exemple au travers de la publicité ou du prix du bien) (Bass, 1969)[24], qui sont autant de paramètres à estimer dans les modèles. Le niveau de ces paramètres influe sur la forme de la courbe de diffusion. Dès lors, au niveau désagrégé, il peut exister plusieurs schémas de diffusion d'une innovation, en fonction des groupes d'individus sélectionnés selon certains critères socio-économiques, tels que l'âge, le revenu, ou le niveau d'éducation par exemple.
En supposant que le produit innovant concerne l'accès et l'utilisation d'un service en ligne, une fracture numérique peut avoir lieu lorsqu'un groupe d'individus (par exemple les plus diplômés) a un rythme d'adoption soutenu dès les premières périodes, alors qu'un autre groupe (les moins diplômés par exemple) adopte le site à un rythme plus lent. Dans ce cas, les deux courbes de diffusion démarrent à une période identique, mais possèdent une forme différente. A l'extrême, un groupe peut stagner à un taux d'adoption donné, alors que l'autre peut avoir achevé le processus de diffusion, conduisant à un taux d'adoption proche de 100 %. Un autre type de fracture numérique existe lorsque chaque groupe possède une courbe de diffusion de forme identique, mais débutant à des périodes différentes. Ce second type de fracture est peut-être moins préoccupant, car finalement le groupe des retardataires rejoindra le groupe des premiers adoptants. Mais il se peut que les deux formes de fracture apparaissent en même temps, la

[24] Bass (1969) rassemble les deux facteurs à la fois.

première étant temporelle (les courbes de diffusion - de même forme - débutent à des périodes différentes pour chaque groupe d'individus), la seconde étant structurelle (les courbes de diffusion sont de formes différentes d'un groupe à un autre).
Discerner les sources potentielles de ces fractures, donc les mécanismes d'adoption et de diffusion d'un service en ligne, est une véritable question de recherche qui pourrait selon nous solliciter une série de modèles alternatifs (par exemple les modèles d'agents en interaction), appliqués à la diffusion des innovations (Antonelli, 1997 ; Dalle, 1997 ; Arthur, 1989 ; Deroïan, 2002, Suire, 2002). Les données de navigation pourraient dans ce sens permettre de valider certains de ces modèles.

Bibliographie :

Abdulla, G., (1998), '*Analysis and modelling of World Wide Web traffic*', PhD dissertation, Computer Science, Virginia Polytechnic Institute, May, 125 p.

Adamic, L. A., (2000), 'Zipf, Power-laws, and Pareto - a ranking tutorial', *Working Paper*, Xerox Palo Alto Research Center, http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html

Adamic, L. A., Huberman, B. A., (2000), 'The nature of markets in the World Wide Web', *Quarterly Journal of Electronic Commerce*, Vol. 1, N° 1, pp. 5-12.

Adamic, L. A., Huberman, B. A., (2001), 'The web's hidden order', *Communications of the ACM*, Vol. 44, N° 9, September, pp. 55-60.

Adamic, L. A., Huberman, B. A., (2002), 'Zipf's law and the Internet', *Glottometrics*, Vol. 3, pp. 143-150.

Almeida, V., Bestravos, A., Crovella, M., Oliveira, A., (1996), 'Characterizing reference locality in the WWW', *Proceedings of PDIS'96*, The IEEE conference on Parallel and distributed information systems, Miami Beach, Florida, December, 12 p., http://www.cs.bu.edu/groups/brownbag/abstracts/jan30-97.html

Anderson, R., Bikson, T., Law, S., Mitchell, B., (1995), 'Universal Access to E-mail: Feasibility and Societal Implications', *RAND Report*, N° MR-650-MF, Santa Monica, CA: RAND Corporation, 267 p., http://www.rand.org/publications/MR/MR650

Antonelli, C., (1997), 'The economics of path-dependence in industrial organization', *International Journal of Industrial Organization*, Vol. 15, Issue 6, pp. 643-675.

Arlitt, M., (2000), 'Characterizing web user sessions', *Working Paper*, Internet and Mobile Systems Laboratory, HP Laboratories, Palo Alto, 7 p., http://kkant.ccwebhost.com/PAWS2000/paper_23.pdf

Arthur, W. B., (1989), 'Competing technologies, increasing returns, and lock-in by historical events', *The Economic Journal*, Vol. 99, Issue 394, March, pp. 116-131.

Barford, P., Bestravos, A., Bradley, A., Crovella, M., (1998), 'Changes in Web client access patterns: characteristics and caching implication', Boston University CS Technical Report 98-023, December 4, http://www.cs.bu.edu/~best/crs/cs591/S99/Schedule.html

Bass, F. M., (1969), 'A new product growth model for consumer durables', *Management Science*, Vol. 15, pp. 215-227.

Beaudouin, V., Licoppe, C., (2002), 'Parcours sur Internet', *Réseaux*, Ouvrage collectif, Vol. 20, N° 116, 314 p.

Bucklin, R. E., Lattin, J. M., Ansari, A., Bell, D., Coupey, E., Gupta, S., Little, J. D. C., Mela, C., Montgomery, A., Steckel, J., (2002), 'Choice and the Internet : from clickstream to research stream', *Marketing Letters*, Vol. 13, n° 3, pp. 245-258.

Bucklin, R. E., Sismeiro, C., (2003), 'A model of web site browsing behaviour estimated on clickstream data', *Journal of Marketing Research*, Vol XL, August, pp. 249-267.

Catledge, L., Pitkow, J., (1995), 'Characterizing browsing in the World Wide Web', *Computer Networks and ISDN Systems*, Proceedings of the Third International World Wide Web conference, Vol. 27, Issue 6, pp. 1065-1073.

Christ, M., Krishnan, R., Nagin, D., Kraut, R., Günther, O., (2001), 'Trajectories of individual www usage : implications for electronic commerce', *34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, Vol. 7, 8 p.

Cordoba, J. C., (2001), 'Zipf's law : a case against scale economies', Mimeo, University of Rochester.

Crovella, M., Taqqu, M., Bestravos, A., (1998), 'Heavy-Tailed probability distributions in the World Wide Web', in Adler, Feldman, Taqqu (Eds), Birkhauser, *Applications of Heavy-Tailed probability distributions*, , Boston, pp. 3-25.

Cunha, C. R., Bestavros, C. A., Crovella, M E., (1995), 'Characteristics of WWW client-based traces', *Working Paper*, Boston, MA, Computer Science Department, Boston University, http://cs-www.bu.edu/faculty/crovella/paper-archive/TR-95-010/paper.html

Curien, N., Gensollen, M., (1989), '*Prévision de la demande de télécommunications : Méthodes et modèles*', Eyrolles, Paris, 458 p.
Dalle, J-M., (1997), 'Heterogeneity versus externalities in technological competition: a tale of possible technological landscapes', *Journal of Evolutionary Economics*, Vol. 7, pp. 395-413.
De Palma, A., Droesbeke, C., Lefèvre, C., (1991), '*Modèles de diffusion en marketing*', Paris, PUF, Collection Gestion, 126 p.
Deroïan, F., (2002), 'Formation of social networks and diffusion of innovations', *Research Policy*, Vol. 31, Issue 5, pp. 835-846.
DiMaggio, P., Hargittai, E., (2001), 'From the digital divide to digital inequality : studying Internet use as penetration increases', *Working Paper*, Center for Arts and Cultural Policy Studies, Princeton University, N° 15, 25 p.
DiMaggio, P., Hargittai, E., Celeste, C., Shafer, S., (2004), 'From Unequal Access to Differentiated Use: A Literature Review and Agenda for Research on Digital Inequality', in Kathryn Neckerman (Eds), *Social Inequality*, New York: Russell Sage Foundation, 73 p., http://www.princeton.edu/~eszter
Fildes, R., Kumar, V., (2002), 'Telecommunications demand forecasting : a review', *International Journal of Forecasting*, Vol. 18, Issue 4, October-December, pp.14 -45.
Forman, C., (2003), 'The corporate digital divide', *Working Paper*, Graduate School of Industrial Administration, GSIA, Carnegie Mellon, August, 33 p.
Gabaix, X., (1999), 'Zipf's law for cities: an explanation', *The Quarterly Journal of Economics*, August, Vol. CXIV, Issue 3, pp.739-767.
George, E., (2004), 'La fracture numérique en question', *in* E. Guichard, *Mesures de l'Internet*, Les Canadiens en Europe (Eds), pp. 152-165.
Glassman, S., (1994), 'A caching relay for the World Wide Web', *Working Paper*, Systems Research Center, Digital Equipment Corporation, Palo Alto, 10 p.
Goldfarb, A., (2002a), 'Understanding Internet clickstream data', *Working Paper*, Joseph L. Rotman School of Management, University of Toronto, March, 30 p., http://www.rotman.utoronto.ca/~agoldfarb
Goldfarb, A., (2002b), 'Analyzing Website Choice using Clickstream Data', in Michael R. Baye. (Eds), *Advances in Applied Microeconomics*, Vol. 11, The Economics of the Internet and E-commerce, Elsevier Science Ltd, pp. 209-230.
Goldfarb, A., (2003), 'State dependence at Internet portals', *Working paper*, Joseph L. Rotman School of Management, University of Toronto, July, 47 p., http://www.rotman.utoronto.ca/~agoldfarb
Guichard, E., (2003), 'Does the 'Digital Divid' Exist ?', in P. van Seters, B. Fortman and A. Ruitjer (Eds), *Globalization and its new divides: malcontents, recipes, and reforms*, Amsterdam, Dutch University Press.
Guichard, E., (2004), '*Mesures de l'Internet*', Ouvrage collectif sous la direction d'Eric Guichard, Les Canadiens en Europe (Eds), 309 p.
Hargittai, E., (2002), 'Second-Level digital divide. Differences in people's online skills', *First Monday*, *Peer-Reviewed Journal on the Internet*, http://www.firstmonday.dk/issues/issue7_4/hargittai
Hargittai, E., (2003), '*How Wide a Web? Inequalities in Accessing Information Online*', Ph.D. , Sociology, Princeton University.
Huberman, B. A., (2001), '*The laws of the web. Patterns in the ecology information*', MIT Press, 128 p.
Kling, R., (1998), 'Technological and Social Access on Computing, Information and Communication Technologies', *White Paper for Presidential Advisory Committee on High Performance Computing and Communications*, Information Technology, and the Next Generation Internet, July, http://www.slis.indiana.edu/faculty/kling/pubs/NGI.htm
Le Guel, F., (2004), '*Analyse économique du comportement des internautes. Mesure, adoption et usages*', Thèse de l'Université de Rennes 1, 297 p.
Le Guel, F., Pénard, T., Suire, R., (2004), 'Une double fracture numérique', in E. Guichard (Eds), *Mesures de l'Internet*, Les Canadiens en Europe, pp. 115-125.

Le Guel, F., Pénard, T., Suire, R., (2005), 'Adoption et usage marchand de l'Internet : une étude économétrique sur données bretonnes', *Economie et Prévision*, A paraître.
Lebart, L., Beaudouin, V., (2003), 'Données dynamiques et mesures de flux sur Internet', Journée Action Spécifique CNRS, GET/ENST, France Telecom R&D, 17 septembre, ENST Paris, http://www.enst.fr/egsh/AS_fluxinternet
Mahajan, V., Peterson, R. A., (1985), 'Models for innovation diffusion', *Sage University Paper Series on Quantitative Applications Social Sciences*, N° 07-048, Newbury Park, CA: Sage, 87 p.
Mansfield, E., (1961), 'Technical change and the rate of imitation', *Econometrica*, Vol. 29, Issue 4, pp. 741-766.
Mitzenmacher, M., (2003)., 'A Brief History of Generative Models for Power Law and Lognormal Distributions', *Internet Mathematics*, Vol. 1, Issue 2, 17 p.
Montagnier, P., Muller, E., Vickery, G., (2002), 'The digital divide: diffusion and use of ICTs', *Report*, Information, Computer and Communications Policy Division, Directorate for Science, Technology and Industry, OECD, 77 p.
Montgomery, A. L., Faloutsos, C., (2000)., 'Trends and patterns of WWW browsing behavior', *Working Paper*, Carnegie Mellon University, 18 p, http://www.andrew.cmu.edu/user/alm3
Montgomery, A. L., Li, S., Srinivasan, K., Liechty, J. C., (2004), 'Modelling online browsing and path analysis using clickstream data', *Working Paper*, GSIA, #2003-E26, September, 36 p, http://www.andrew.cmu.edu/user/alm3
Montgomery, A. L., Li, S., Srinivasan, K., Liechty, J. C., (2004), 'Modelling online browsing and path analysis using clickstream data', *Working Paper*, GSIA, #2003-E26, September, 36 p, http://www.andrew.cmu.edu/user/alm3
NTIA., National Telecommunications and Information Administration., (1995), 'Falling through the net: a survey of the have nots in rural and urban america', *U.S. Department of Commerce Report*, July, http://www.ntia.doc.gov/ntiahome/fallingthru.html
OCDE., (2001a), 'Understanding the digital divide', *OECD Publications*, Paris, 32 p
Pareto, V., (1896), '*Cours d'Economie Politique'*, Genève: Droz.
Park, Y-H., Fader, P. S., (2004), 'Modelling Browsing Behavior at Multiple Web Sites', *Working paper*, Wharton School, University of Pennsylvania, 51 p., http://www-marketing.wharton.upenn.edu/people/faculty/fader.html
Pitkow, J. E., (1998), 'Summary of WWW characterizations', *Computer and ISDN Systems*, Vol. 30, Issue 1-7, April, pp. 551-558.
Rallet, A., Rochelandet, F., (2003), 'La "fracture numérique" : une faille sans fondement ?', *XXXIXe Colloque ASRDLF*, Lyon, 1,2,3 septembre 2003, 22 p.
Rice, R. E., Katz, J. E., (2003), 'Comparing Internet and mobile phone usage: digital divides of usage, adoption, and dropouts', *Telecommunications Policy*, Vol. 27., Issues 8-9, September - October, pp. 597-623.
Rogers, E. M., (1995), '*Diffusion of innovations*', London, The Free Press, 4th edition, 246 p.
Simon, H. A., (1955), 'On a class of skew distribution functions', *Biometrika*, Vol. 42, Issue 3/4, December, pp. 425-440.
Suire, R., (2002), '*Réseaux sociaux et géographie économique*', Thèse de l'Université de Rennes 1, 284 p.
Tauscher, L. M., (1996), '*Evaluating history mechanisms: an empirical study of reuse patterns in WWW navigation*', M. Sc. Thesis, Department of Computer Science, University of Calgary, June, 173 p., http://www.cpsc.ucalgary.ca/grouplab/papers/1996/96-Tauscher.Thesis/thesis/entire_thesis.pdf
Urzùa, C. M., (2000), 'A simple and efficient test for Zipf's law', *Economics Letters*, Vol. 66, Issue 3, pp. 257-260.
Zipf, G. K., (1949), '*Human Behavior and the Principle of Least Effort*', Addison-Wesley.