

Stilling our information hunger:

*How can we achieve better search results
and disclose new sources of information
using the Internet*

by **Thomas Clever**

Contents

5 For Your Information

QUESTIONING THE INFORMATION LANDSCAPE

7 Dismantling a Search Engine

THE PRINCIPLE WORKINGS OF GOOGLE

11 User Habits

DYNAMICS OF THE MASSES

13 The Six Handshakes

NETWORK ARCHITECTURE AND THE INTERNET

21 Post-it

THE CURRENT INTERNET LANDSCAPE

23 The (Pre)Search Engine

COINCIDENCE STRIKES THE PREPARED MIND

For Your Information

QUESTIONING THE INFORMATION LANDSCAPE

- 1 Marshall McLuhan, media theorist, coined the idea of the 'global village' that described how mass media collapse space and time barriers in human communication, enabling people to interact on a global scale.
- 2 Weblogs or 'blogs' are personal online diaries on news, general or particular subjects.
- 3 VPRO 2006a
- 4 Rogers 2000, p.25
- 5 Metacrawlers search and combine the results of multiple search engine like Yahoo and Google

In the last twenty years the world has moved towards the idea of the global village.¹ The World Wide Web and the Internet have bridged social gaps and geographical distances. Generations to come will grow up without knowledge of the world prior to the existence of the World Wide Web. Whether it is used for informational purposes or as a social communications device, the young people of today switch between 'virtual' and 'real' world without hesitation. Today's users have no trouble sharing their music, films and keep track of each other's lives through weblogs.²

The World Wide Web has become a place where knowledge no longer has an author. 'Knowledge' used to be something to strive and work hard for but is now handed to us instantly on a 'silicon' platter in the form of Wikipedia – a 'collective memory' – that grows larger everyday. Previously, knowledge was a status symbol; 'knowledge is power'. Google has materialised the right to index and globalise all information. Wikipedia allows anyone to add anything. Information has become a common good to be modified by anyone. The webpages are now all we need to obtain information and acquire knowledge.

The majority of Internet users make use of search engines Yahoo, MSN and Google. For a long time now, these three companies have occupied the top three spots as most popular gateways to access the Internet. Estimations of the number of users of these search engines lay around the 91% mark³, but search engines only index up to 16% of all websites that are online today.⁴ Metacrawlers⁵ improve one's chances of finding more websites but even the best can only reach up to 50%. How is it possible that so many registered websites cannot be accessed through search engines? These initial finding at the start of my research brought up many questions.

The unparalleled popularity of search engines might not be as justified as one might think. Is it possible to get access to the remaining percentage of websites through other channels? Is the information withheld in this 84% not of relevance to one's search query? Maybe it is unfeasible to have access to all corners of the web, but surely, the opportunity to decide the relevance of a website for oneself is of huge importance. Is the unparalleled popularity of search engines justified? Are there alternatives? Is it possible to get access to the remaining 84% of websites through other channels? Is the information withheld in this 84% not of relevance to one's search query? Maybe it is unfeasible to have access to all corners of the web.

I started searching on the Internet to find answers to these questions but soon found out Internet was not going to be satisfactory. The fact my questions were not answered in depth by search engines and websites, formed an answer to my questions in itself. I was not just looking for information but also for 'context', something to help me interpret information and make new connections. I was trying to acquire knowledge in order to answer a complex problem myself.

Are websites, then, at all relevant when trying to acquire this knowledge? Most websites can answer simple questions but in complex cases a broad range sources of information are needed to develop critical and analytical perspectives. Much of the material that comes up in a search engine consists of fragmented pieces of information, synopsis and summaries.

The main question I hope to answer in this thesis is how we can achieve higher quality search results and via which means? Information that can give depth to complex questions and connections to further someone's research. Results that have a form of integrity and authorship so that it can be placed in a context and its roots can be traced back.

If we think of 'knowledge' as no more or less than drawing up connections between different ideas, then finding connections between sources would bring us one step closer to acquiring knowledge and being able to answer questions.

How can we find higher quality sources of information and disclose sources that would otherwise remain untapped when using only a search engine? Will this redefine the principal workings of search engines or how we should use the Internet?

Dismantling a Search Engine

THE PRINCIPLE WORKINGS OF GOOGLE

- 6 Cybernetics is the study of feedback and derived concepts such as communication and control in living organisms, machines and organisations. (Wikipedia 2007a)
- 7 Idea first introduced by Heinz von Foerster, one of the originators of Cybernetics.
- 8 Vint Cerf, Vice President and Chief Internet Evangelist of Google, VPRO 2006b
- 9 Seomoz 2007a

'Googling' has become a verb in everyday language. If something cannot be found on Google – it seems – it cannot be found anywhere. For many of us, Google forms primal access to the Internet and dictates what the Web consists of. It has become the all-knowing information tool – the oracle of Delphi – suggesting it has access to 'all' of the world's online information.

Before defining where Google lacks in the search process, it is important to understand the workings and popularity of Google. Like any good product, Google offers its consumers the things they are looking for. When looking up timetables or the latest news on celebrities, any search engines appears to function perfectly well. With its advanced search options we are able to get even more detailed information. However, I found searching for information on more complex questions and queries give me dissatisfactory results. Results did not show more depth or variation in sources or information than can be found on Wikipedia. Search engines could be fantastic tools in providing information for specific research fields and digging into virgin territory. Applying Cybernetic⁶ theory, we can define search engines as trivial machines that give predictable answers. Trivial questions have answers that already exist, but interesting questions don't have answers yet.⁷ Why do current search engines act as a 'trivial machines'?

Ranking

The secret of Google's search results or 'hits' lies in its Pagerank system. This determines and ranks the importance of a webpage. Pagerank proves influential in the display and hierarchy of search results. The basic idea is as follows: entering a search query, Google will screen all web pages for corresponding information. Google has 'scored' websites for a web pages' importance. The importance score for any web page is expressed in a non-negative real number. This is assigned by the Pagerank algorithm that starts at a random website and then follows links to other websites. After having visited a few websites, the algorithm stops and starts at a new random website. The probability of the algorithm visiting the indexed website via links from other websites determines the 'score'. Thus, the indexed web seems to become a democracy where pages vote for the importance of other pages by means of linking to each other. However, this 'voting' is not as democratic after all.

Google ranks on popularity, which is determined by the number of in- and out links from a website, but also "what is thought to be relevant to the consumer".⁹ However, when Google determines the score, it makes it impossible to judge how relevant the given hits really are. The unfairness of this system lies in the fact that wealthy, powerful websites – often belonging to big corporations – contain more in and out links, thus creating a higher score.

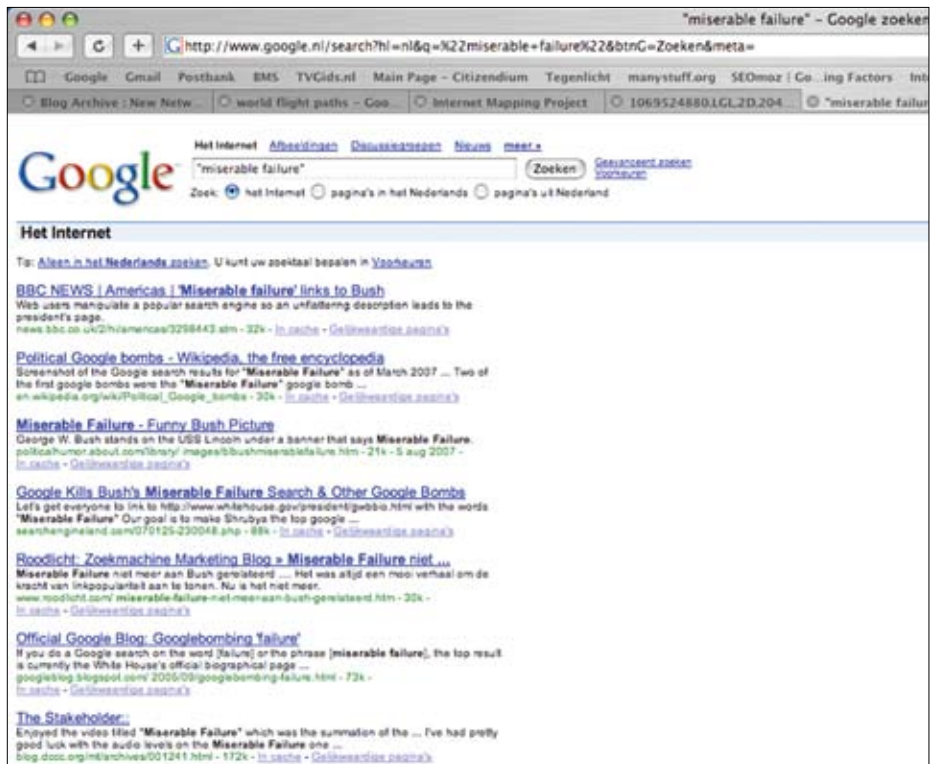
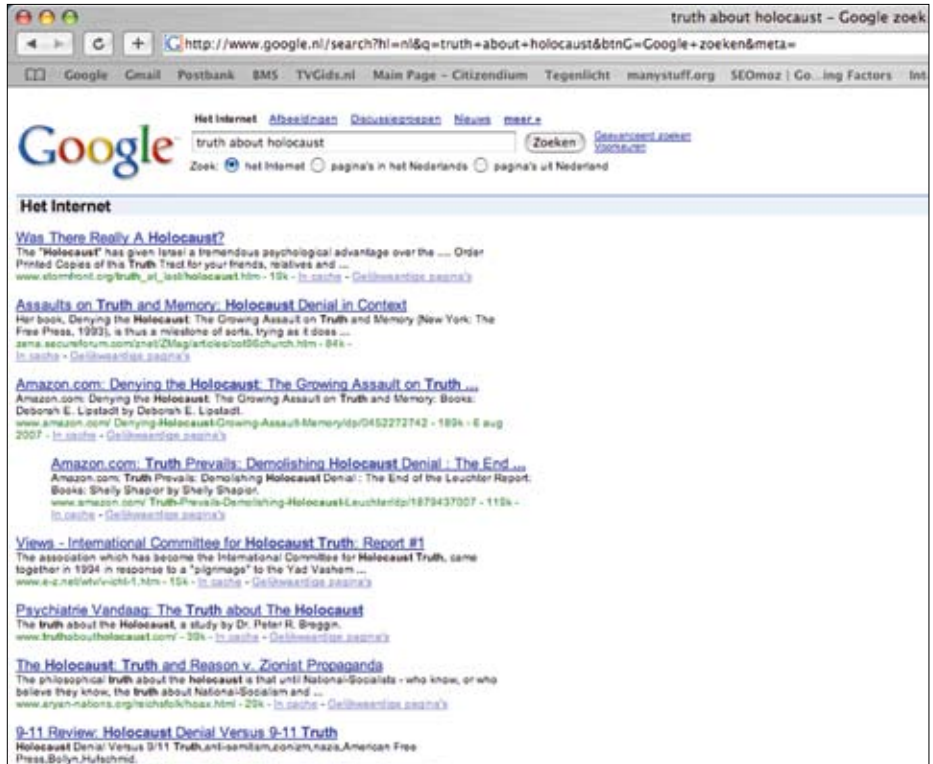
There are controversial factors that determine the hierarchy of Google's search results. For example, Google links to many large corporations and vice versa. The ranking system differentiates between importance and popularity of in- and out links. Since Google has a huge amount of in- and out links, having a Google link is seen as very important and will be ranked much higher than another link to this website. Another factor is that Google gives manual (occasional) authority or manipulation to webpages.⁸

Trivial and politically correct?

Still, Google only indexes 16% of websites. This low percentage is partly due the immense growth of the World Wide Web, causing major delay to the index process of search engines. The other reason is more troubling. Search engines are biased and

Dismantling a search engine

- Over the years some kinks in Google's workings have been discovered. For instance there is Google bombing, famously making George Bush the top result for the search *miserable failure*. In order to change this the algorithm has to be tweaked; editors of the editor-less engine have to step in.
- Another example is that of searches surrounding the terms *Jew* and *Holocaust* have produced anti-Semitic results. The only way to challenge this is to appeal to Google and change their editor-less engine.



10 Lovink 2005, p. 9

11 BBC News 2006a

12 Upstream network traffic flows away from the local computer toward the remote destination. Conversely, downstream traffic flows to the user's computer. Traffic on most networks flows in both upstream and downstream directions simultaneously, and often when data flows in one direction, network protocols often send control instructions (generally invisible to the user) in the opposite direction.

Wikipedia downstream: "Data shows that for the week ending Feb 10, 2007, 70% of Wikipedia's upstream visits came from search engines, with 50% from Google alone. Google's share of Wikipedia's upstream traffic from Google has increased by 19% over the past year, at the same time that Wikipedia's market share of US visits increased by 143%.

If it seems like Google is sending more traffic to Wikipedia than in the past, it's because it is. The percentage of Google's downstream traffic going to Wikipedia increased by 166% year over year (week ending 2/10/07 vs. week ending 2/11/06). Last week Wikipedia was the #3 website in Google's downstream, after Google Image Search and MySpace." (Hitwise 2007a)

13 Vint Cerf, VPRO 2006c

14 VPRO 2006d

15 Masters of Media 2007a

16 Singer 2005, p. 54

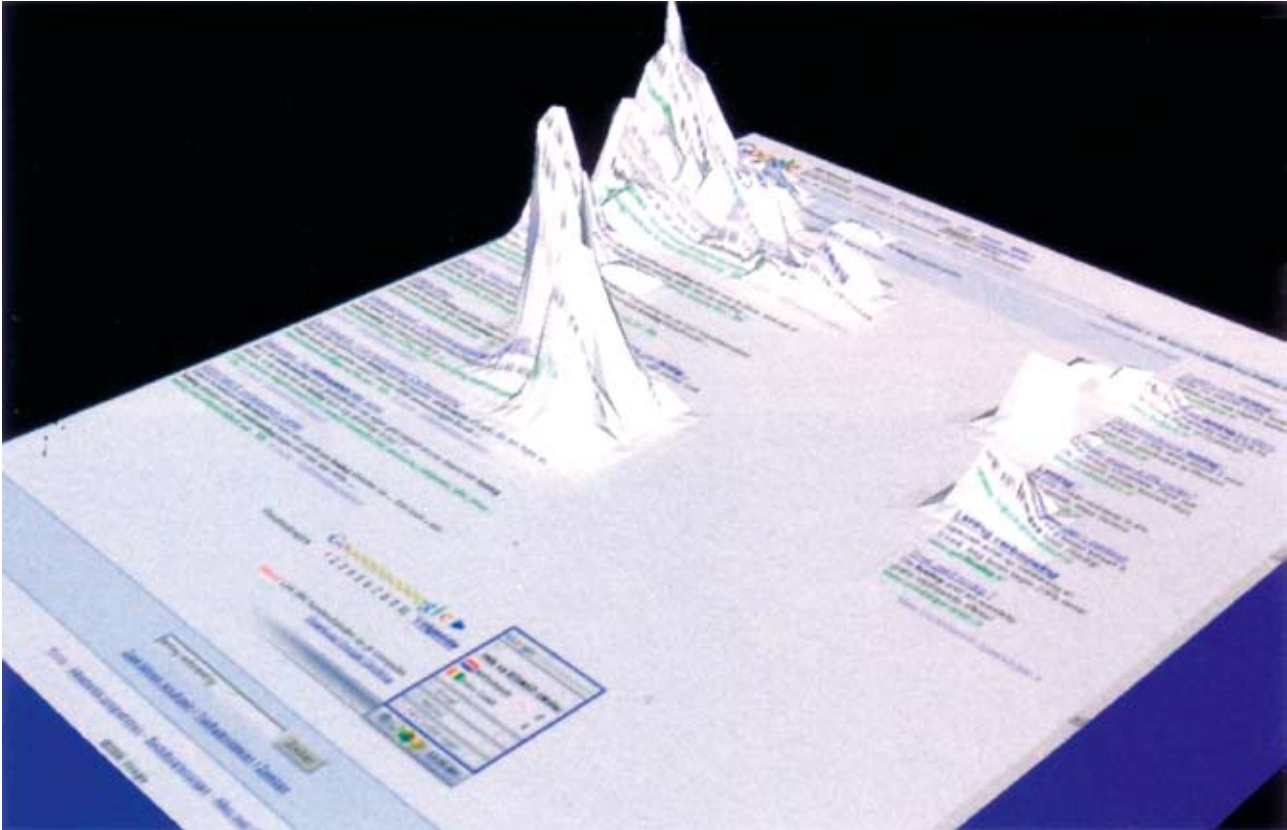
"come with specific built-in social, cultural and aesthetic agendas".¹⁰ These give preference to certain websites over others. Most Internet users are unaware of the design faults of search engines. The Google Corporation claims its dominant role, as player on the 'information market' is by no means biased or influenced by political choices. However, Google chooses its business partners carefully and has recently teamed up with the Chinese government to censor search results that might be politically undesirable.¹¹ It claims these concessions do not stand in relation to the benefits of the Chinese people that now have 'access' to Google. Looking at the Chinese economic growth, one might be inclined to think this deal is more beneficial from an economic point of view. Is Google not biased when Wikipedia entries have become top hits for just about any search term? There must have been an edit to Google's index that changed Wikipedia's trust ranking. Downstream figures certainly show some love affair is going on.¹²

"Google tries to be neutral"¹³ but says it relies heavily on critical thinking on the part of the recipient of information. However, one can question this point of view when so many people see Google as an 'all-knowing' information source or simply do not have the capability of critical thinking (yet).

Google offers the illusion of free service, democracy, precision and objectivity. In fact the searching and surfing habits of its users is what pays Google. Google does not speak of users but consumers. It makes its money by advertising, which is why it is profitable for Google to direct users to certain websites. Google records the search movement of the individual user. Over time it adjusts its search results to the 'benefit' of the user, based on these data. What better way to market things when you already know where the interests of the users lie?

In current light, Google's slogan 'Don't be evil'¹⁴ might seem rather ironic. Its indexing methods leave much to be desired. Google is determined to provide universal access to 'all' of the world's information. Or to put it in the words of Sergey Brin, founder of Google: "It would be like the Mind of God".¹⁵ Even though the comparison with the mind of God is highly speculative, the comparison does translate Google's ambitions to have total control over the world's information. In his theory of history, Karl Marx said: "man cannot be free when he is subject to forces that determine his thoughts, ideas and even his nature".¹⁶ Thoughts and ideas - in the form of search results - are not free, they are pre-programmed and market driven. As we will come to see, users appear to be limited by ideas and nature in the way these search engines are used.

User habits



User Habits

DYNAMICS OF THE MASSES

17 John Maeda is professor at MIT's Media Lab.

18 Maeda 2006, p. 91

19 Van Helsdingen 2007, p. 64-65

20 Buchanan 2003, p. 120

- 'Eyetracking' diagram of person searching for information using Google. Peaks reflect the places the user looked at. The higher the peaks, the longer the viewing time.
- Eye movement of the user. Larger size dots reflect longer viewing time. In total the viewing time was no longer than 1.1 sec.

As said, 85% of Internet users use Google to access the Internet. It is not surprising that Google's dominancy is kept alive with such staggering percentages. What is it that makes Google so popular? According to John Maeda,¹⁷ Google's power lies in its simplicity and it's simple search experience. "Think of the power of Google which runs from a simple, lightweight text input box in your web browser. (...) More appears less by simply moving it far, far away".¹⁸ Studies show users look no longer than 1.1 sec. at the results presented to them.¹⁹ The top three links are the most 'clicked' links and it is very unlikely the average users will look at the results beyond the first page i.e. results #11+ in Google). Also it seems to be hard to differentiate between Google's results and their own sponsored links. Even though Maeda's point about its simplicity is a valid one, apparently after the search has been performed the user is given little support on where to go next. The performed search also shows the user the database has retrieved over a million hits within 0.21 sec. A job well done, one might say. Looking at it from the opposite point of view, one could also question whether Google has really done its job since one might still need to go through a million links to find the correct information. Again, Google proves to be political in ordering and presenting the information in this way.

Why do 85% of Internet users (including myself) use Google? The explanation lies for a large part in our nature and can be explained on the basis of two sociological phenomena: *Information cascades* and *Groupthink*.

Information cascades work on the basis of 'actors' basing their actions on previously observed actions of others, and making the same choice independent of one's individually preconceived ideas. Therefore creating a limited 'space' for all the actors to move about. For example: the bus stops in front of my flat every 5-7 minutes. However, on Sundays when the university is closed (my house in on the route to the university) the bus only comes 3 times per hour. After 3 years I have still not taken the trouble to find out at what time my bus arrives on Sundays. Instead, I look out of the window to see if people stand at the bus stop. The more people there are, the more likely it is I will start making my way towards the bus stop. I have information that tells me that the bus turns up 3 times an hour and so do all the other people. I assume one of them knows at what time the bus arrives, as do all the other people that live in the neighbourhood. So to supplement my private, information I turn to others for more information since I assume one of them must know the timetable by heart. The problem starts when people make the decision to move to the bus stop in sequential order. If the first person has the wrong timetable information the sequence that comes after this will be based on false conceptions. The fundamental problem here lies in the case that people stop relying on their private information. The more important a decision, the less likely a cascade will hold. The likelihood to continue looking out of my window for the bus or, for that matter, use Google to access the Internet is more likely to hold.

As website owner, getting a website indexed is probably the most crucial part of 'surviving' online. An exclusion from major search engines diminishes the chances of being found. Psychologist Irving Janus explored the way groups come to make decisions and presented his findings in the idea of *Groupthink*. This idea concluded: "group dynamics limit the group's ability to legitimately consider alternative options."²⁰ In this case, the creator of a website has the freedom to choose which sites to link to but 'group dynamics' has triggered every one to link to Google. People's fear of exclusion and group dynamics is partly the cause for Google's monopoly position and keep it firmly in place.

21 Sunstein 2006, p. 25

22 Edward Bernays is seen as one of the founding fathers of public relations. Bernays was one of the first to attempt to manipulate public opinion using the psychology of the subconscious. (Wikipedia 2007b)

23 Bernays 1930, p. 2

Groupthink and *Information cascades* showed how group dynamics causes people to collectively embrace Google. The so-called Jury Theorem²¹ invalidates these ideas to some extent and shows human collectives are not only 'herd animals' but also can be smart in mobs. Its principle justification lies in the demonstration that groups are likely to do better than individuals if majority rule is used and each person is more likely to be correct. Despite its simplistic reasoning it sense. One is more likely to get a correct answer when asking a group of specialist on a specific subject instead of randomly selected people. Under these conditions, the majority can be trusted. As it turns out, the chance of a group being right improves when:

- people were unaffected by whether their votes would be decisive;
- people would not be affected by one another's votes;
- and the probability that one group member would be right would be statistically unrelated to the probability that another group member would be right.

Ask a group of people to determine how many jellybeans are in a glass jar and the majority will call a number that will be close to the correct amount. Here the majority rule proves to be effective. Asking the same group of people how many atoms are held in the glass jar, the majority will fail to come to the right amount. Topics such based on general knowledge or feelings, like polls do, generate useful statistics when asked to any group of people. However some questions acquire preconceived knowledge and therefore asking (a group of) specialists will acquire a far more accurate answer, hence a 'knowledge cluster'.

Last but not least, another problem appears once the user has arrived at the location of the search engine. Entering the correct search query is crucial to finding 'satisfactory' results. The users form the first hurdle in achieving our desired results when we formulate our query when we visit a search engine. Users are forced to formulate search queries in one or two words instead of a sentence with ending with a question mark – like in everyday speech. Even with Google's advanced search options, it does not function like a real life conversation. In a real life conversation questions are posed and discussed; a process that cannot be captured in two words or seconds.

To conclude with the words of Edward Bernays²²: "The conscious and intelligent manipulation of the organised habits and opinions of the masses is an important element in democratic society. Those who manipulate this unseen mechanism of society constitute an invisible government, which is the true ruling power of our country."²³

The Six Handshakes

NETWORK ARCHITECTURE AND THE INTERNET

24 Gladwell 2000, p. 61

25 Milgram 1967, p. 60-67

26 At the time, it seemed less likely that black and white people operate in the same social circles.

At the root of network theory, we find other explanations for the development of Google's popularity and search engines alike.

"Fred Jones of Peoria, sitting in a sidewalk cafe in Tunisia, and needing a light for his cigarette, asks the man at the next table for a match. They fall into conversation; the stranger is an Englishman who, it turns out, spent several months in Detroit. 'I know it's a foolish question; says Jones, 'but do you by any chance know a fellow named Ben Arkadian? He's an old friend of mine, manages a chain of supermarkets in Detroit...'

'Arkadian...Arkadian...' the Englishman mutters. 'Why, upon my soul, I believe I do! Small chap, very energetic, raised merry hell with the factory over a shipment of defective bottle caps'

'No kidding!' Jones exclaims, amazed.

'Good lord, it's a small world, isn't it?'"²⁴

Network theory finds its roots in a paper published by social psychologist Stanley Milgram.²⁵ He was one of the first to start an investigation into so-called 'The Small World' phenomenon, also known as the 'six degrees of separation'. Milgram tried to come to a better understanding behind the structure of social networks, and in his experiment he presented fair evidence that the world is in fact smaller than one might think.

Handing out 160 letters to random people in New York, Milgram made a first attempt into tracing social routes. All letters had to reach a person in Boston, address not specified, only a name. The letters that arrived came through the hands of only three of the person's friends. More strikingly, almost all letters arrived in only six steps. In a second phase, Milgram made it more difficult for the letters to arrive. He incorporated the idea of racial segregation of white and black America into the experiment.²⁶ Letters were handed out to random white people in Los Angeles and had to arrive to randomly selected black people in New York. Again, most letters reached their destination within six steps.

The concept of six degrees of separation might seem a little far-fetched and not be more than a curiosity. However, Milgram's experiment made a great contribution to a new kind of science in mathematics and physics: graph and network theory, which forms the basis of the architecture of the Internet, the organisation of the World Wide Web and the operation of search engines.

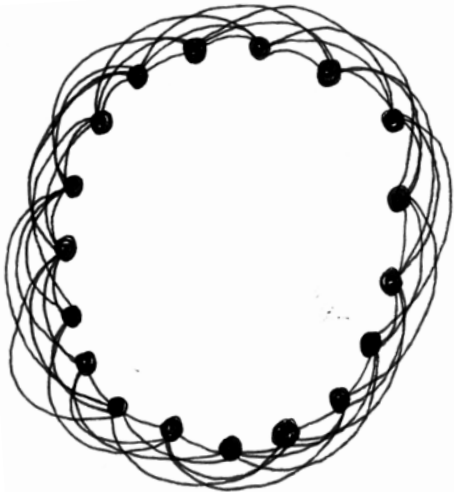
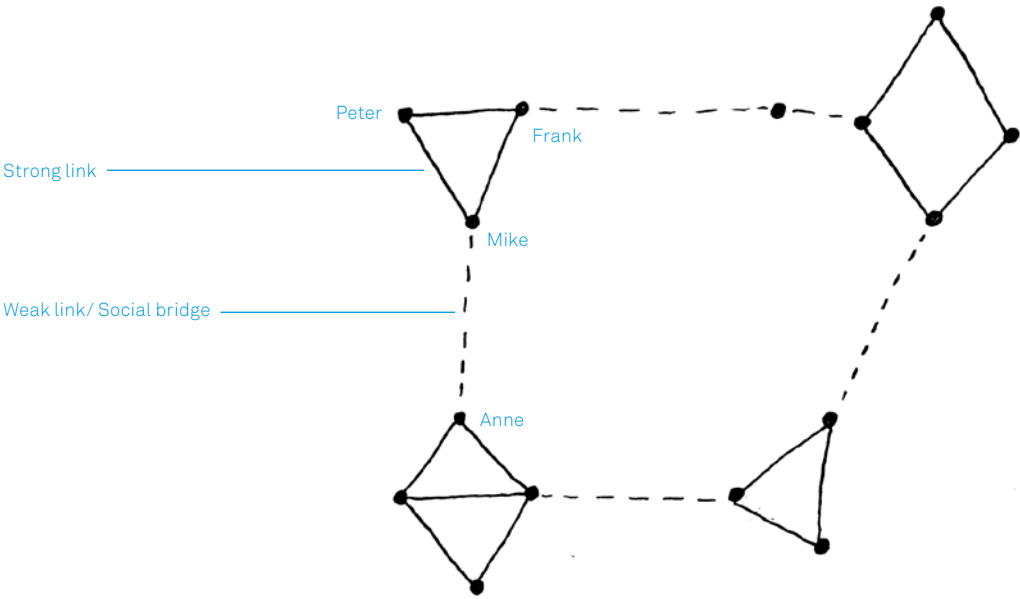
Graph's and Networks

Let assume a person – Frank – has 50 acquaintances and take the number for the world population to be 6 billion. How quickly can Frank be linked to everyone in this world?

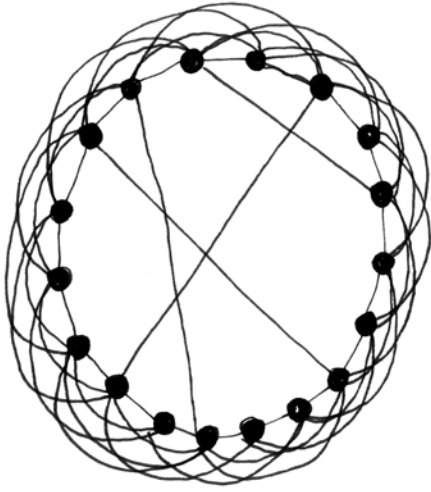
If Frank has 50 people he is linked to and those 50 are linked to 50 others we have already reached a number of 2500 people out of 6 billion within two degrees of separation. Six degrees of separation will give us a number larger than 15 billion, two and a half times the world's population!

$$1 > 50 > 2500 > 125,000 > 6,250,000 > 312,500,000 > 15,625,000,000$$

The six handshakes



..



...

27 Granovetter 1973, p. 1360-1380

- Social clusters; tied together by weak links.
- Ordered network. Each point is connected to its three nearest neighbours.
- Ordered network with random links added. The random links shorten the path from one point to a point on the otherside of the network.

The principle of weak links

In theory Frank knows the entire world's population within six steps. It would be too easy to say everyone on this planet is linked within six degrees of separation. The great difficulty with this idea lies with Frank and his friends. Many of Frank's 50 friends are likely to be socially tied between them. This concept is known as 'clustering'. People are not linked randomly all over the world. The social network of Frank, and any social network for that matter, shows overlaps between friends. A real social network will not grow quite as fast as the simple calculation indicates. Suppose Frank has a best friend called Mike who has a colleague called Anne. The link between Frank and Mike is 'strong' but the link between Frank and Anne is 'weak' because Frank can only be linked to her through Mike. Since Frank and Mike are best friends, it is likely they share a common friend: Peter. Suppose we would remove one of the strong links between Frank, Mike and Peter. This would hardly have any effect on the degrees of separation between the other two. In general, strong links appear in triangles, therefore, changing one of the links does not have a devastating effect within the cluster and 'social distances'. One might think strong links are crucial in holding a social network together, but sociologist Mark Granovetter first introduced the importance of weak links²⁷ between people as social bridges within social networks.

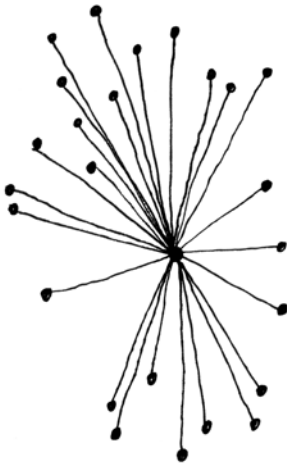
Anne and Mike were linked through work. They are colleagues but no more than that. When we knock out the link between Frank and Anne – i.e. Mike – Frank loses touch with Anne. The degrees of separation would drastically increase for Frank to get to Anne again and possible access to her social network.

This simple example gives us one fundamental insight: weak links are of crucial importance in linking between communities. Without these, social groups would form isolated cliques. Connecting the ideas of 'the strength of weak links' and Milgram's experiment present a better insight into the world of complex networks, such as the Internet and the World Wide Web.

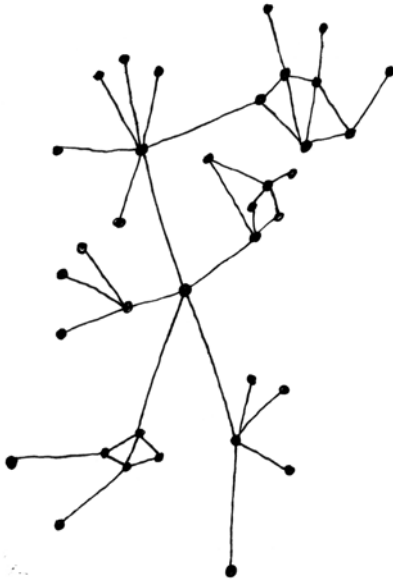
Inspired by Milgrams' experiment, mathematicians Duncan Watts and Steve Strogatz worked out that networks give rise to clusters and cliques, in the way that it happens in social networks. Ordered networks do not behave conform the small world property, i.e., the six degrees of separation. Random networks, on the other hand, make for small worlds but do not form clustering. A social network had to be a peculiar mix of randomness and order! To explore these ideas further Watts and Strogatz conducted hundreds experiments and developed computer models to illustrate the following idea. Suppose a number of dots, 1000, each connected to their 10 nearest neighbours. This would amount to about 5000 links. Following, a number of random links were added to the model. The model showed a mixture of dots and random links that was somewhere between ordered and random. The initial network that had been created had the potential of 45 links running between 10 neighbours. Watts and Strogatz came up with the so-called 'clustering coefficient'. In the model, 2 out of every 3 dots were linked. The clustering coefficient is obtained dividing the number of actual links (2) in the network over the total number of possible links (3). Therefore $2/3$, or 0.67 became the networks clustering coefficient (C.C.). The coefficient tells us how closely knit a network is. For example, a social network with a C.C. close to 1.0 means all friends are good friends with each other.

Adding more disorder, of the initial 5000 links, 1% (50) random links were added to the ordered network. While having little impact on the clustering coefficient – the number dropped from 0.67 to 0.65 – the added random links did have an amazing influence on the degrees of separation between the dots. Without random links, the networks' degrees of separation were around 50, including random links the number plummeted to 7!

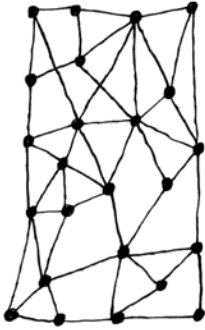
The six handshakes



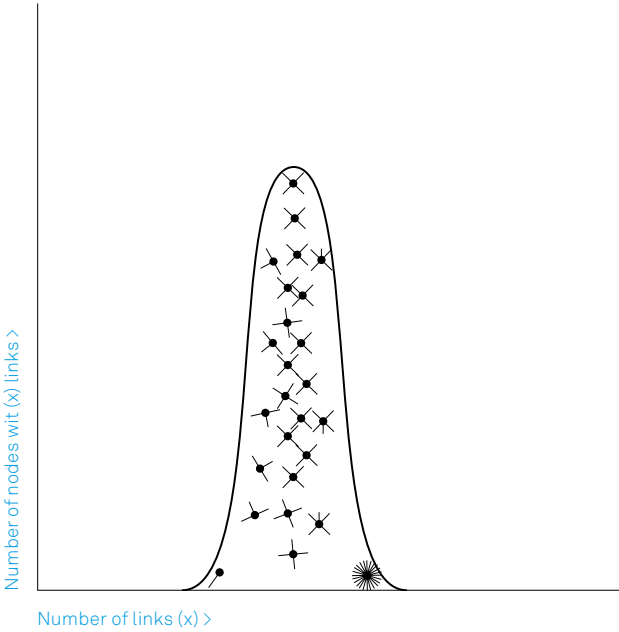
•



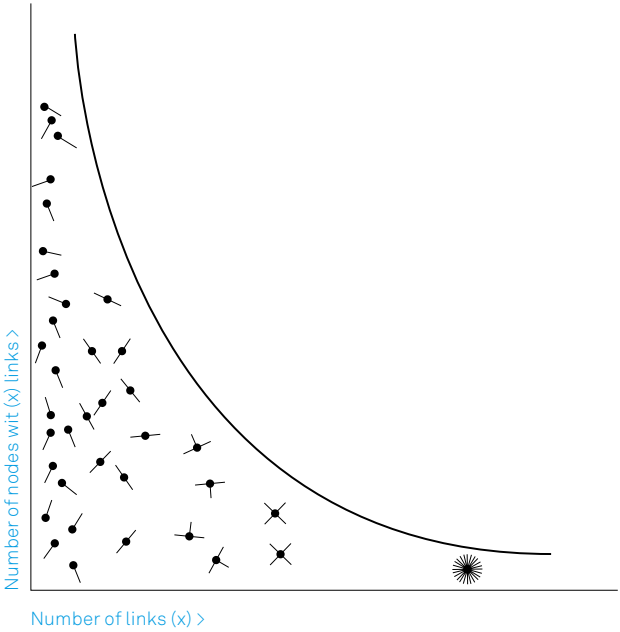
••



•••



••••



•••••

The six handshakes

- 28 In mathematic theory called the *Power Law*, in economics in is called the *80/20 Rule*.
- *Centralised network*
 - *Decentralised network*
 - *Distributed network – envisioned structure for the Internet*
 - *The Bell curve shows the distribution of nodes with x amount of links. Such a curve applies to egalitarian network.*
 - *The Power Law: curve reflects an aristocratic network. The least amount of nodes possess the greatest amount of links.*

Taking the number of the world population the degrees of separation in an ordered network would be something like sixty million. The computer model showed that adding even as little as 3 random links for every 10,000 dots, the degrees of separation dropped to 5. Even a very small number of weak links or random links have an enormous influence on the degrees of separation in a social network. Clustering does not stop at the boundary of social networks. It shows that one can be very local in choosing one's friends as long as a small fraction of the population has some long distance friends.

The Internet Movie Database provides possibly the largest set of data for a social network (in this case of actors). The average degrees of separation between the actors were about 3.65 with a clustering coefficient of 0.79. The database holds 225,226 actors that are linked to 61 other actors on average. The acting world is both highly clustered and a small world at the same time.

A real world network that holds the same properties as the IMDB case are power grids or road maps. In these grids each element is roughly linked to three others. Some junctions are highly clustered, yet the degrees of separation remain low. These examples show that real world networks can be highly clustered and small world all at once. Despite its simplicity, it is of immense importance to the architectural foundation of the Internet.

The birth of the Internet

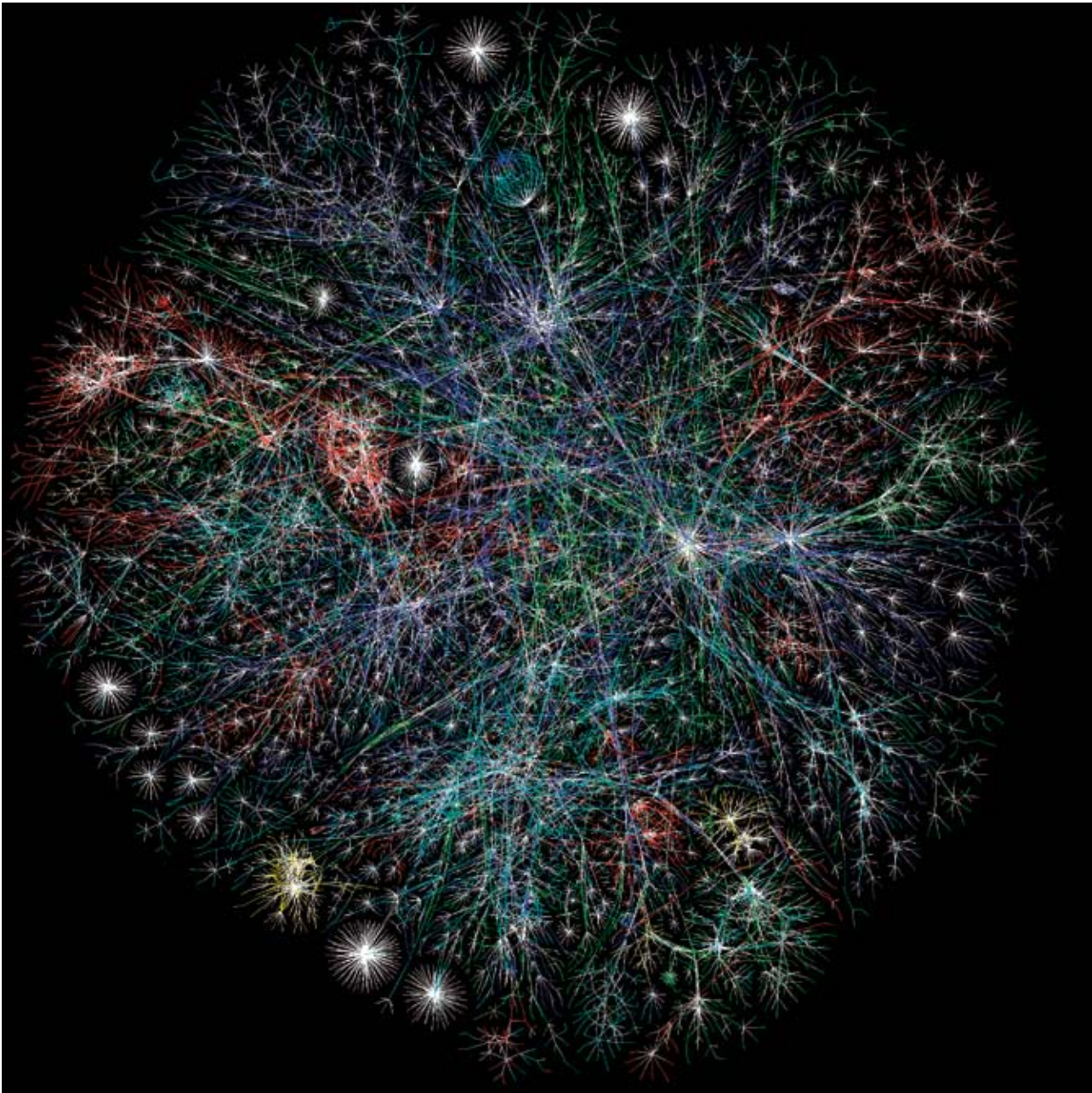
The Internet came into existence during the cold war era, after the Cuban missile crisis. Then known as ARPANET, the communications network was specially designed to withstand a nuclear missile attack and sustain minimal damage. The initial design of ARPANET knew two types of distributed networks. One had close similarities with the ordered networks; the other has a fishnet type structure. In the latter case, attacking a random point would not devastate the network like it would when a hub were to be struck.

However, over the years the Internet has developed into a network that does show the characteristics of a small world. Test and studies have shown that email traffic from Hong Kong to Helsinki goes through no more than four degrees of separation. Even computers that seem impossible to link to each other never needed more than ten degrees of separation. Paradoxically, the clustering of computers is far greater than would be expected from a random network; neither does it show the characteristics of the original model of the Internet. Instead, it's a self-organised small world network, which needs very few steps to get from one computer to another.

Strangely enough, the Internet has no central authority that traffics information. The Internet grows with unparalleled speed but it is not built by adding random links. It must therefore be a different kind of small world network.

As the models of Watts and Strogatz showed, highly clustered networks can easily be connected through a few random links. Grannovetter added the idea of the strength of weak links. However, their 'egalitarian' networks did not feature the element of time. The small world networks of Watts and Strogatz started with ten thousand points and after adding a few random links the network was still the same size containing ten thousand elements. Time is a crucial element in the development of networks such as the Internet.

The architectural structure of Internet envisioned each node in the network to have roughly an equal amount of links. However, the Internet of today shows a network structure full of hierarchy: hubs with many links. This growth might seem completely uncontrollable since anyone can start a webpage and publish it online. However, the growth of the Internet can be explained according to a rule known as the *Power Law*.²⁸ This law states: each time the number of links double the number of nodes with this



30 Buchanan 2003, p. 86

31 Buchanan 2003, p. 130

- [Map of the Internet showing IP-addresses linking to each other. The image clearly shows it is no longer a distributed network.](#)

amount of links becomes about five times less. The picture of the Internet might appear to be chaotic and random, in fact it is subject to this kind of order principle.

What this tells us is that one is far more likely to find a website with a high number of links rather than a website with a low amount of links. In the words of mathematician Barabasi: “The probability of finding a document with a large number of links is rather significant, the network connectivity being dominated by highly-connected web pages. (...) The probability of finding a very popular address to which a large number of other documents point is non-negligible and an indication of the flocking sociology of the World Wide Web”.³⁰

In the instance of *Groupthink* and the *Power Law*, we have come to understand why search engines such as Google have grown to be so dominant. They form the hubs of the World Wide Web. The *Power Law* gives legitimacy to hubs and to the rise of aristocratic models of networks. In such a network it is inevitable that over a period of time ‘the rich get richer’. Hubs tend to dominate network activity and the way networks expand.

One final striking feature in the development of hubs in networks is the difference between ‘ethereal’ and real world networks. The Internet is an ethereal network. There are no limits in how many links a website can have. In the case of physical networks, the Power Law phenomenon is rare. Like the electrical grid, curiously enough the world airport network is one of the egalitarian kind. Even though large airports are referred to as hubs, any two airports in the world can be linked through no more than 5 steps. “As the network grows, the rich get richer for a time, and hubs do emerge. But eventually the most highly linked elements begin to lose their advantage in gathering new links.”³¹ When limitations or costs come into play, the rich stop getting richer and the network moves towards the egalitarian kind. In the case of airports congestion and the limitations of building new airstrips caused smaller airports to catch up and taking over flight routes.

Post-it

THE CURRENT INTERNET LANDSCAPE

- 32 O'Reilly 2006
- 33 Google statistics show the most popular search queries of 2006 to be nearly all Web 2.0 ented:
 - 1 – bebo
 - 2 – myspace
 - 3 – world cup
 - 4 – metacafe
 - 5 – radioblog
 - 6 – wikipedia
 - 7 – video
 - 8 – rebelde
 - 9 – mininova
 - 10 – wiki
- 34 Lovink 2007, p. 3
- 35 Sunstein 2006, p. 186
- 36 NRC Handelsblad 20 July 2007, p.13
- The Information Architects Japan published a map of the most popular websites to date. The 'weather forecast' reflects the predicted future development of individual websites in relation to Web (x).(x).

After the Internet bubble exploded (2001) a new generation of web applications emerged, known as Web 2.0. Web 2.0 typifies itself by user-added content and the main principle of the web acting a platform.³² The success stories of Web 2.0 include the likes of Flickr, YouTube, Wikipedia and weblogs or 'blogs'. Such companies currently dictate new developments on the World Wide Web.

People are given more control over content on these platforms and can therefore be considered democratic areas on the Web. In the case of blogs, it shows people are interested in material of an opinionated and subjective nature.³³ In the case of Myspace and Facebook users are able to landscape their social network. Linking to friends and friends of friends, these networks can be seen versions Milgram's small world idea.

Unfortunately, the way these networks and blogs are used at the moment, these remain platforms full of highly speculative material.

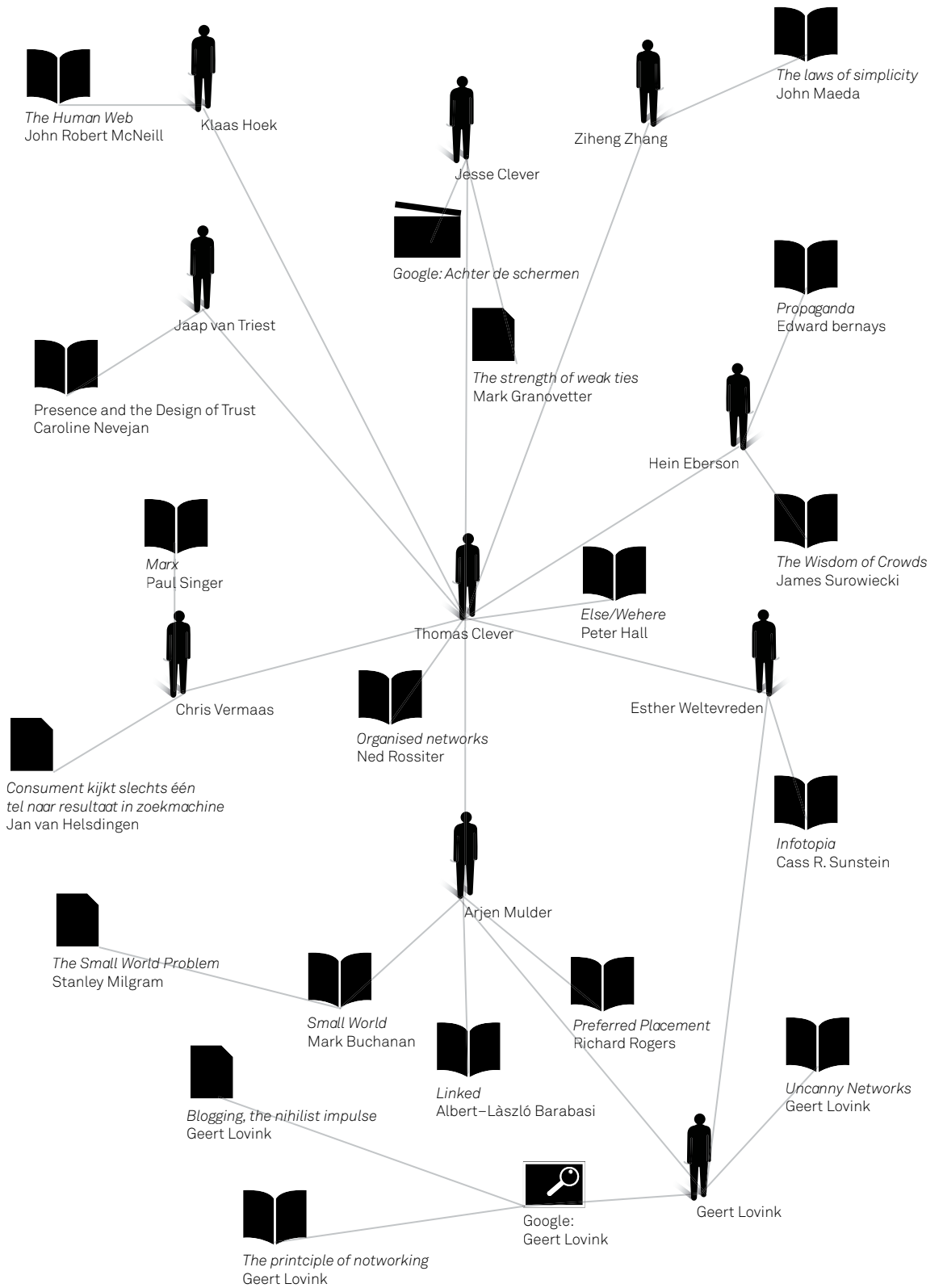
Blogs come closer into the direction of delivering new sources of information and offer the possibility to get into contact with the author or 'blogger'. Blogs are great for reviewing trends and some have grown to be highly informative on new developments in areas like technology. Good blogs are updated regularly and often comment of current affairs, therefore reflecting the sentiment of a specific time and place.

However, the downside of these blogs lies mainly in the active bloggers themselves. Most blogs do not concentrate on one specific topic. It is nearly impossible to say anything about their general content and categorise them. Blogs are often a collection of fragmented, informal thoughts on subjects, often not based on knowledge but on ideas or feelings of the blogger. "What ordinary blogs create is a dense cloud of 'impressions' around a topic."³⁴

Some blogs are popular, with an active group of people commenting on 'posts'. The downside to this is that discussions on the platform leads to polarisation of the group and inaccurate answers to subjects discussed. The Jury Theorem does not hold in these cases. With blogs, the reader is influenced by all the other comments and does therefore not hold the same quality as a one-on-one conversation. "Even the best blogs lack anything like prepublication peer review, and their speed and informality often ensure glibness, superficiality, confusion and blatant error."³⁵ Even though this is part of what makes it fun to read the diaries, generally blogs cannot be seen as information sources of high quality or integrity or function as a search engine.

Google is also part of the Web 2.0 era and works hard as becoming a platform to cater for all types of information. Alternative options for search engines to combat Google's popularity receive little publicity. The only serious development at the moment is that of Quaero (a French government initiative) and Theseus (a German government initiative).³⁶ Other models are often labelled utopian, devalued, since these do not operate on the market place.

The (Pre)Search Engine



The (Pre)Search Engine

COINCIDENCE STRIKES THE PREPARED MIND

- Diagram shows the channels through which I received my main sources of information. Most of the sources were recommended by my social network while discussing my research with them.

During the process of working out an answer to the main question of this thesis, I have gained insight in the problems of current search engines and why these prove unsatisfactory. It has led me to come up with a proposal on how and with what means to search in order to answer my questions.

Research Experience

The majority of information sources I used were not found through search engines. I discovered most of these through my network of friends, teachers and acquaintances. The ties to these people proved to disclose new sources of information that had otherwise remained unknown to me. People I contacted about my research did not answer my questions, but instead helped formulate reasoning and deepened my knowledge during discussions. Often sources would in turn link to new information or insights. My social network functioned as a kind of search engine in itself. Entering words in a search field would not have given me the same quality of search results that I was able to extract from my social surroundings.

Analysing my own research experience has given me insight in how a research process develops. With the risk of stating the obvious, doing extensive research and attempting to find answers to complex questions still requires one to investigate sources, visit libraries and talk to specialists. Many people only use websites as primary sources of information for research, a popular phenomenon among high school students and even university undergraduates. Using search engines will find you 'known' answers, but excludes the possibility of finding the unknown. Research requires sources that can always be traced and consulted afterwards. Most results that come up with search engines are useless since the author is unknown or information given seems to run a hidden agenda. Online information has become hybrid since the authors are not visible most of the time.

Democracy

As we have seen, the Internet and the World Wide Web are subject to the *Power Law*: only a handful of points (like Google) in the network possess the greatest number of links. On the other hand, there is the egalitarian model of networks in which all points in the network have roughly the same amount of links; the architectural structure originally envisioned for the Internet. Clearly, the workings of current search engines are questionable. Their ability to access only 16% of the websites has much to do with their own wrongdoing. With the unfairness of their systems and the limited understanding of the user, gaining access to all corners of the web remains difficult. But surely, the opportunity to decide the relevance of a source of information for oneself is of huge importance.

In theory, to have democratic access to information and obtain knowledge through a search engine, its network architecture should be like an egalitarian network. A structurally democratic network model for searching – free from hubs – in which every user would have *individual* control over the search process. Such an engine should not take commercial success into account. Instead it should develop according to initiatives like the Unix operating system or Camino browser. A cooperative group of people exchanging information, knowledge and debate, merely out of 'ideological' convictions. In a sense this is where real democracy lies: discussion and debate. Only then are we to obtain better search results. As in the case of airports, real world networks tend to even out hub-like behavior. A more democratic search engine would most likely develop into this direction if it were to be linked to a real world network.

- 37 For example: network theory is a subject that covers a range of specialist fields such as biology, traffic management, physics, mathematics, information design, medicine, geology, etc. Current search engines can find information on these topics but do not lay connections between different fields. To understand network theory, I found it useful to talk to geologists, biologist and designers to view this from different perspectives and find new approaches.

Obtaining better search results

Searching through a social network proved to be fruitful. Taking the six degrees of separation, clustering and the random links from the Watts and Strogatz models, a ‘real world’ search engine should be more like a database of people, categorised according to specialism.

Random links showed to bring down the degrees of separation drastically and took Milgram’s six-handshake theory from utopian idea to a realistic possibility. Six handshakes away from anyone in the world would put us – theoretically speaking – in reach of any desired piece of information or at least a human connection. The links that tie individuals together in a network should be meaningful rather than being random. Instead of linking people merely by their social connections, individuals within the network could be rearranged according to their specialism or field expertise. Activating our ‘weak’ links to navigate through communities, we can contact individuals and consult their *personal database* of sources. Providing a platform where one is able to contact different specialists on a specific subject offers the opportunity to sharpen questions, ideas and find a variety of new sources of information in the process. Different types of sources from different fields of expertise hold relation to each other. Finding these connections proved impossible to find through a list of links.³⁷

The Jury Theorem justified asking specific questions to a group of specialists. Therefore, basing a search engine around the idea of clustering people on knowledge would generate higher quality search results. Hence, individual people as search results. Of course knowledge clusters already exist on some level like students in the same course or a geological journal. But individuals are often not in contact with similar clusters on the other side of the globe and outsiders do not have access to these clusters. Internet transcends our slowly changing information cocoons; geographical boundaries, school communities and workplaces. Those who work together closely are likely to see things in the same way. Consequently, setting up ‘knowledge clusters’ of specialists from around the globe would enhance the chances of receiving a variety of sources. The Internet enables us to step out of the confinements of the social circle and step into a cluster of specialists that are not influenced by the same social and political forces.

This search engine proposal could be best accessed prior to using existing search engines. It would host the possibility of contacting specialists through one’s social or knowledge network in order to consult (them or their) sources of information. In turn the specialists can refer to other people, information or give insights. Existing search engine become useful to find out where to, buy a book, but are too often used as a starting points in researching. and do not offer the possibility of discussion about ideas or material. Also, searching with Google excludes the possibility of coming across something unexpected. Google gives what the user wants. *Serendipity* is based one the idea that sometimes one accidentally discovers something fortunate, while looking for something else entirely. By interacting with others we run this ‘risk’ of finding something unexpected; new connections.

Disclosing better sources of information demands one to get into contact with other people. Still, a real world network offers more solutions to one’s research than a ethereal networks like current search engines.

Concluding along the lines of Cybernetics: finding high quality information and disclosing new sources should be done through a *non-trivial machine*. A machine with unpredictable outcomes that caters for the development of research and novel ideas instead of looking for (known) answers. The answer to *this* (still) lies in humans and human interaction.

Books

- Albert-László Barabasi, *Linked*, London: PLUME/Penguin Group 2003
- Cass R. Sunstein, *Infotopia*, New York: Oxford University Press 2006
- Edward Bernays, *Propaganda*, New York: Horace Liveright 1930
- Geert Lovink, *Uncanny Networks, dialogues with the virtual intelligentsia*, Cambridge/Massachusetts: MIT Press 2004
- Geert Lovink, *The principle of notworking, concepts in critical Internet Culture*, Amsterdam: HvA Publicaties 2005
- James Surowiecki, *The wisdom of crowds*, New York: Anchor Books 2004/2005
- Janet Abrams, Peter Hall, *Else/Where: Mapping, new cartographies of networks and territories*, Minneapolis: University of Minnesota Design Institute 2006
- John Maeda, *The laws of simplicity*, Cambridge/Massachusetts: MIT Press 2006
- Mark Buchanan, *Small World*, London: Phoenix 2002
- Ned Rossiter, (2006), *Organised Networks - Media Theory, Creative Labour, New institutions*, Rotterdam: NAI Publishers
- Paul Singer, *Marx*, Rotterdam: Lemniscaat 1980
- Richard Rogers, *Preferred Placement, knowledge politics on the web*, Maastricht: Jan van Eyck Akademie 2000

Articles

- Geert Lovink, 'Blogging, the nihilist impulse', *Eurozine*, 1 February 2007, p. 1-16
- Malcolm Gladwell, 'Six degrees of Lois Weisberg', *The New Yorker*, January 11, 2000, p. 52-63
- Mark Granovetter, "The Strength of Weak Ties"; *American Journal of Sociology*, Vol. 78, No. 6., May 1973, p. 1360-1380
- Stanley Milgram, 'The Small World Problem', *Psychology Today*, May 1967, p. 60-67
- Jan van Helsdingen, 'Consument kijkt slechts één tel naar resultaat in zoekmachine', *Adforesult*, March 2007, p. 64-65
- O'Reilly 2006
- Tim O'Reilly, 'What is Web 2.0, Design patterns and business models for the next generation of software', 9 May 2005, 5 August, 2007,
< <http://oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> >

Websites

BBC News 2007a

news.bbc.co.uk, 'Google censors itself for China', 25 January 2006, 5 August, 2007
<news.bbc.co.uk/1/hi/technology /4645596.stm>

Master of Media 2007

Michael Stevenson, 'New Network Theory - Siva Vaidhyanathan' mastersofmedia.hum.uva.nl, Weblog entry, Masters of Media, 28 June 2007, 5 August 2007
< mastersofmedia.hum.uva.nl/2007/06/28/new-network-theory-siva-vaidhyanathan/#more-615 >

www.univie.ac.at/constructivism/HvF.htm – Heinz Von Foerster

www.pangaro.com/published/cyber-macmillan.html – Cybernetics

semoz 2007

www.semoz.org, 2 April 2007, 5 August 2007,
< www.seomoz.org/article/search-ranking-factors >

VPRO 2006a

www.vpro.nl, VPRO, 5 August 2007,
< www.vpro.nl/info/tegenlicht/webspecial/ >

wikipedia 2007a

"Cybernetics", Wikipedia: The Free Encyclopedia, 5 August 2007,
< en.wikipedia.org/wiki/Edward_Bernays >

wikipedia 2007b

"Edward Bernays", Wikipedia: The Free Encyclopedia, 5 August 2007,
< en.wikipedia.org/wiki/Cybernetics >

Other

VPRO 2006b, VPRO 2006c, VPRO 2006d

'Google: Achter het scherm', Tegenlicht, VPRO Nederland 3, 7 May 2006

Images

p.8 – < www.google.nl >

p.10 – Jan van Helsdingen, 'Consument kijkt slechts één tel naar resultaat in zoekmachine', *Adforesult*, March 2007, p. 64-65

p.18 – < www.opte.org >

p.20 – < www.informationarchitects.jp >

Stilling our information hunger:
*How can we achieve better search results
and disclose new sources of information
using the Internet*

by [Thomas Clever](#)

Master Thesis | MaHKU | Editorial Design 2007

Graphic design
[Thomas Clever](#)

Printing
[Printnet.be](#)

I would like to thank the following people for their
support, advice and enthusiasm:

[Arjen Mulder](#)

[Chris Vermaas](#)

[Esther Weltevreden](#)

[Gert Franke](#)

[Hein Eberson](#)

[Henk Slager](#)

[Jaap van Triest](#)

[Jan van Toorn](#)

[Jelle Henze](#)

[Lude Franke](#)

[Fam. Clever](#)