

# Network Knowledge: Gleaning new knowledge from the Internet

---

## AN EXAMPLE OF FORECASTING WITH SOCIAL NETWORK ANALYSIS

Author(s):

Chris Sadler, School of Computing Science at Middlesex University, [c.sadler@mdx.ac.uk](mailto:c.sadler@mdx.ac.uk)

London, July 2008



**Leonardo da Vinci**



This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

# Contents

---

<b>Abstract</b> .....	<b>3</b>
<b>Introduction</b> .....	<b>4</b>
<b>Analysing networks</b> .....	<b>7</b>
1. Measures of Centrality .....	8
<b>Applying Network Analysis</b> .....	<b>11</b>
<b>Conclusion</b> .....	<b>12</b>
<b>Bibliography</b> .....	<b>13</b>

# Abstract

---

Mankind has always lived in communities – the family, the settlement, the tribe, the nation. The communications systems developed in the nineteenth and twentieth centuries, from postal services through the telephone to the internet, allow individuals to belong to virtual communities which can be ephemeral both in time and space.

In spite of this, the age-old community dynamics (the relationships between lesser and more influential or authoritative group members) which can be modelled using social network analysis, still apply to the virtual social networks that we form on the World Wide Web. By these means it may be possible to predict public preferences and shifts in public opinion throughout our complex global world. This potential is of great interest to governments and corporate business everywhere, and the abstract, analytical approach adopted here may be more acceptable to individuals and human rights groups than the alternative, large-scale capture of personal data that these bodies currently employ.

## **Keywords**

Social Network Analysis, Scale-free networks, Betweenness, Forecasting behaviour

# Introduction

---

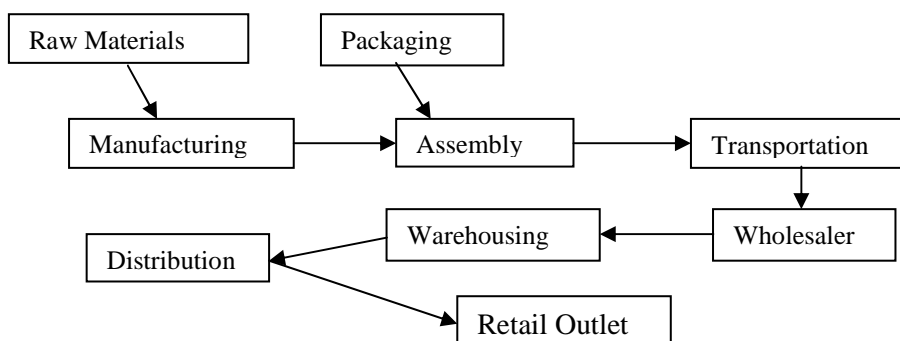
What is a network? At an abstract, graphical level it is a collection of points (called **nodes**) some of which may be joined together by **links**. In the real world, by defining what the nodes actually are and what the links mean, we can use the idea of a network to understand and analyze certain types of complex systems.

Many of the complex systems we want to analyze are related to the development and evolution of human society. Transportation networks are a good example. As soon as communities began to develop and grow, it was necessary to create trading routes between these communities. Here each community (for example, a city state) is a node of the network, and the routes between node are obviously the links. A famous example of a land trade route would be the ‘Silk Road’ which connected Changan in Northern China to Constantinople in the West via a long series of links joining many ancient cities in between (see [The Silk Road: Linking Europe and Asia Through Trade](#)). A famous sea trade route was the ‘Spice Route’ along which cargoes from Goa and Colombo made their way to Lisbon and Amsterdam respectively (see [Spices and herbs go way back](#)).

During the time of the Roman Empire, roads were constructed all over Europe mainly to facilitate the movement of troops. These formed an extensive network and some of them still exist in the modern-day road network, even though new methods of road-building are used and many new roads have been added to the network. The development of railways saw a new transportation network being laid down across countries whilst air-travel has created another (inter-continental) network. In all these networks, the towns and cities (or their stations and airports) form the nodes and the routes themselves provide the links.

Likewise, communications networks have been developed over time, starting with postal systems (based on the existing transportation networks) through telegraph and telephone systems (where the links were copper wires) to broadcasting and satellite systems where the nodes are transmitting and receiving stations and the links are neither physical nor permanent. Similarly there are utility distribution systems which pipe power and water into our homes.

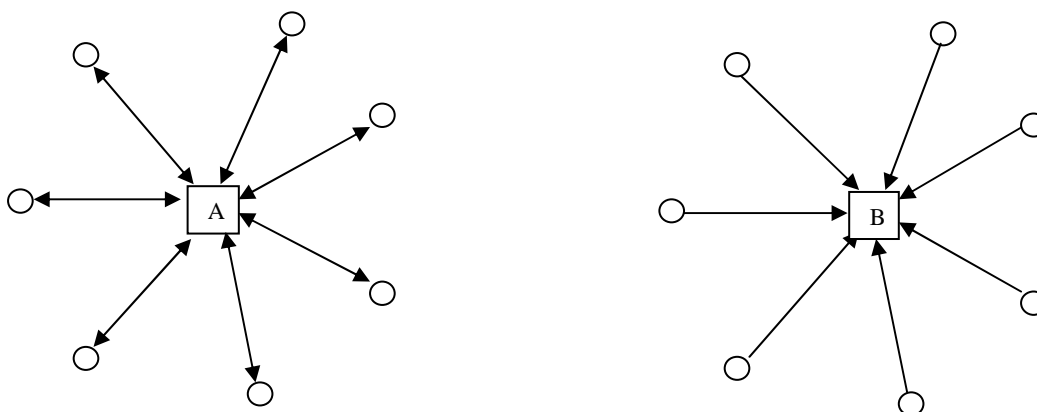
1st Figure: A simple Supply Chain



There are other complex aspects of our society that can usefully be modeled using networks, including social systems themselves where the nodes can be organizations, or groups or even individual people and the links consist of transactions of some sort between them. When we speak of a company’s *supply chain* ([Chang and Makatsoris](#)) we mean the network of companies that exchanges goods and services to provide the commodities that the company sells (see 1st Figure). When we talk of somebody’s *circle of acquaintance* we mean the network of all the people known to (linked to) that person. On the other hand, when we talk about a celebrity’s *fan base*

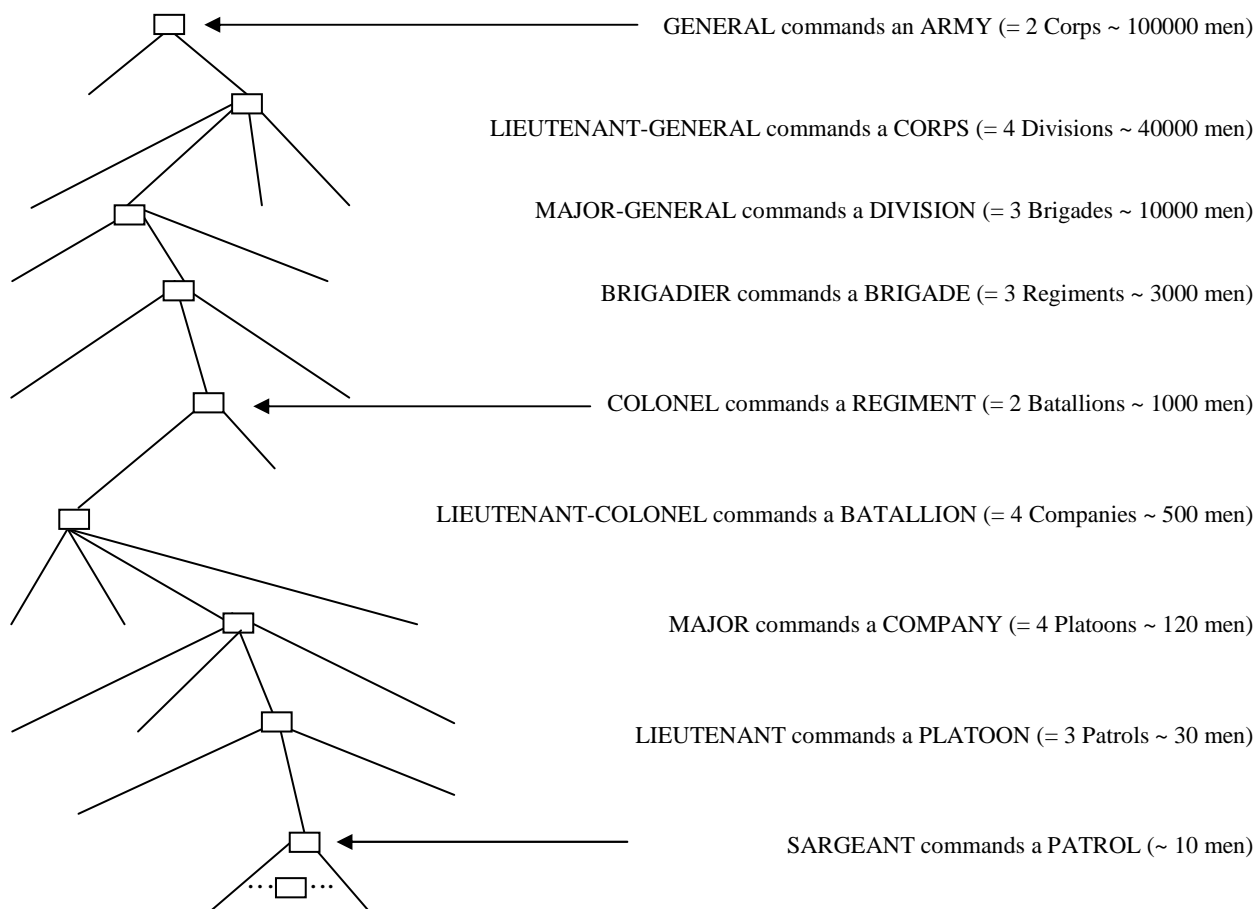
we mean the network of all the people that know (and admire) him or her – the nodes are the same but the nature of the links is different (see 2nd Figure).

2nd Figure: People knowing people: A’s circle of acquaintance and B’s fan base



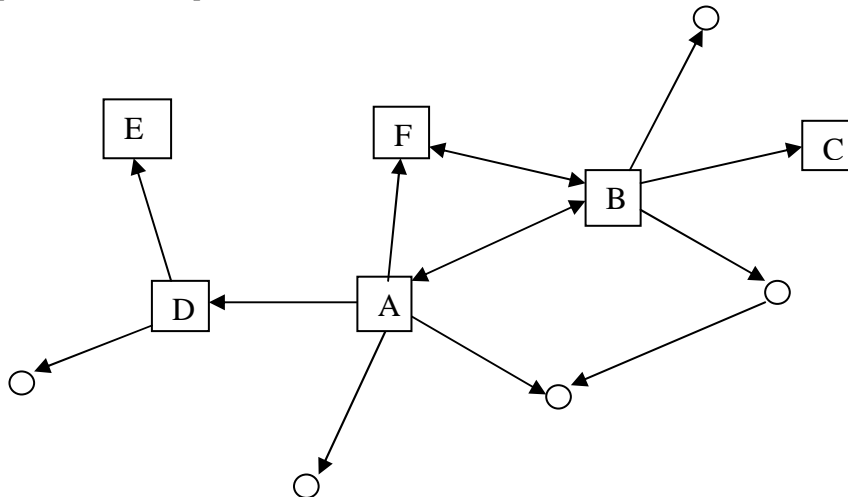
There are many types of networks where the number or nature of the nodes is limited or controlled in some way. The ‘chain’ and ‘circle’ (usually known as a **star** network) discussed above are two examples. Another type commonly found in social networks is the **hierarchy** which characterizes the line of command found in military and many other organizations (see 3rd Figure).

3rd Figure: Typical British Army command structure (see [British and US Military Ranks](#))



All the above networks show actual or intended relationships between the various nodes, but in most cases they do not capture anything that is particularly complicated. In situations where there are no, or few restrictions on which or how many nodes can be linked together, more complicated networks can arise. Consider the example of the circle of acquaintance in 2nd Figure. If we tried to document the acquaintances of some of A's acquaintances, the diagram might look something like 4th Figure. As new nodes are introduced and new links made, the picture becomes more complicated. In 2nd Figure, every person was at most two links away from every other person.

4th Figure: Acquaintances of acquaintances



In 4th Figure, B knows F directly (rather than indirectly through A) but is three links away from E (B -> A -> D -> E). C and E are separated by 4 links. Networks like these can get very complicated very quickly (who has only 3 acquaintances like D?) and the nodes can become highly interlinked<sup>1</sup> so we need to develop more powerful methods of understanding and managing networks than just sketching them out.

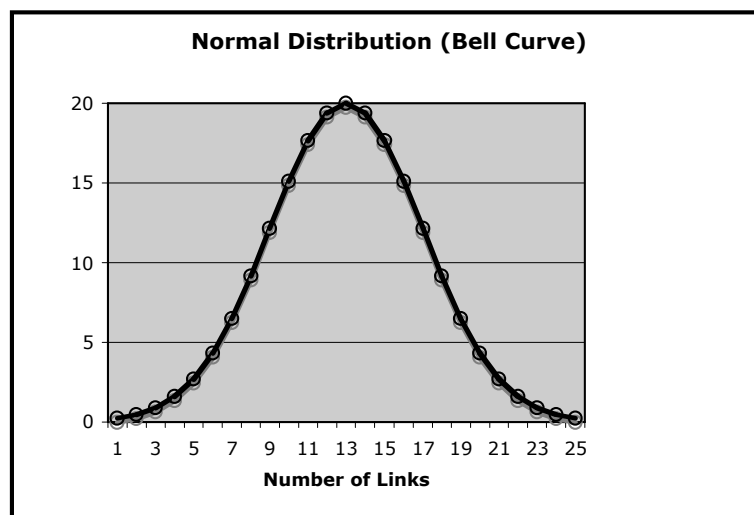
<sup>1</sup> In spite of this degree of complication, it is conjectured that everyone in the world is separated from every other person by as little as six links. This is known as the *Small World* phenomenon.

# Analysing networks

---

When scientists thought about trying to model networks mathematically, they initially assumed that the pattern of links would be random. Thinking about social networks at the level of an individual person, this is clearly not true. If you move into a new city, the people that you link up with are likely to be very specific – new neighbours; new workmates; people at the sports hall or wherever you go for entertainment. But if you consider this *statistically* – a number of people moving to a new city will move into an arbitrary neighbourhood, be working at an arbitrary workplace and seeking arbitrary forms of entertainment. Thus the **random network** is not an unreasonable model when considering the network as a whole. Counting the number of links possessed by each node leads to a graph similar to 5th Figure.

5th Figure: Normal distribution

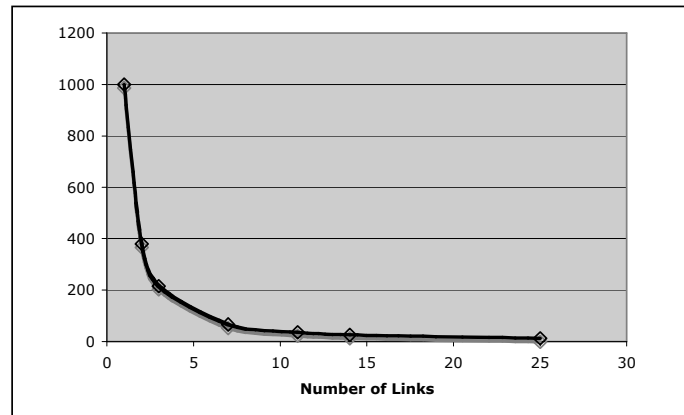


In the normal distribution (the ‘bell’ curve), the nodes are symmetrically distributed about the average value, and, depending on how steep the sides of the bell are, the average value can be considered as the number of links a *typical* node has. Some networks in the physical and social worlds conform to the random network model and some useful results can come from this analysis.

However, some networks, especially certain social networks, exhibit behaviours which are extremely difficult to explain in terms of a random network. One example of this is the way that a rumour can rapidly spread through a community or the speed with which a contagious disease can infect a population. The random model predicts that most nodes are linked to close to the average number of other nodes so that the rumour has to cross many links to get from one end of the network to the other. However, this arrangement cannot account for the speed with which everybody finds out about rumour so soon after it starts. Another phenomenon observed in some networks (especially communications networks) is the apparent robustness against failure – the network keeps functioning even though many nodes have failed. A random linkage predicts that such a network would fracture into several sub-networks at a much earlier stage.

Then Barabasi, Jeong and Albert (Barabasi & Bonabeau, 2003) discovered a new type of network in 1998 when investigating the network of interconnected pages on the World Wide Web. Here, the vast majority of pages had only a few links (less than four) whilst a tiny fraction had thousands of links. This distribution (see 6th Figure) is very different from the ‘typical’ node predicted by random network analysis.

6th Figure: Power Law Distribution



As new nodes are added to the network, they are most likely to link to the most popular nodes (e.g. Google or Yahoo on the Internet) rather than to some arbitrary neighbouring node. This implies that the characteristic curve shown in Figure 6 (the *power law* curve) stays pretty much the same shape no matter how big the network grows. Such networks are known as **scale-free** networks. Since this discovery, the scale-free property has been identified in many types of networks including certain types of social networks. For example, many scientific research papers are written by teams of scientists collaborating together. When the co-authors of all the papers in a particular discipline are linked together, they form a scale-free network. Likewise, the co-stars in Hollywood movies form a scale-free network. More negatively, the victims of an outbreak of a sexually transmitted disease (STD) form a scale-free network.

## I. Measures of Centrality

The key feature of scale-free networks is the existence of those highly-connected nodes, known as **hubs**. This explains the rapid diffusion of a rumour – the rumour goes to all the contacts of the person who starts it. The chances are that at least one of these is a hub, who is probably linked to the other hubs who pass it on more-or-less directly to everybody else. Within the scientific discipline, there are a number of key (hub) professors who are the big stars of their disciplines and everybody wants to conduct research with them.<sup>2</sup> In the Hollywood network, surprisingly, it is not the big box-office stars who are hubs, but the highly talented ‘character’ or supporting actors whom the directors seek to hire because they can rely on them to hold the movie together no matter how temperamental the stars get.

Networks **degrade** when links are broken or when nodes cease to function. Under these circumstances, scale-free networks have shown themselves to be remarkably resilient in the sense that information continues to

---

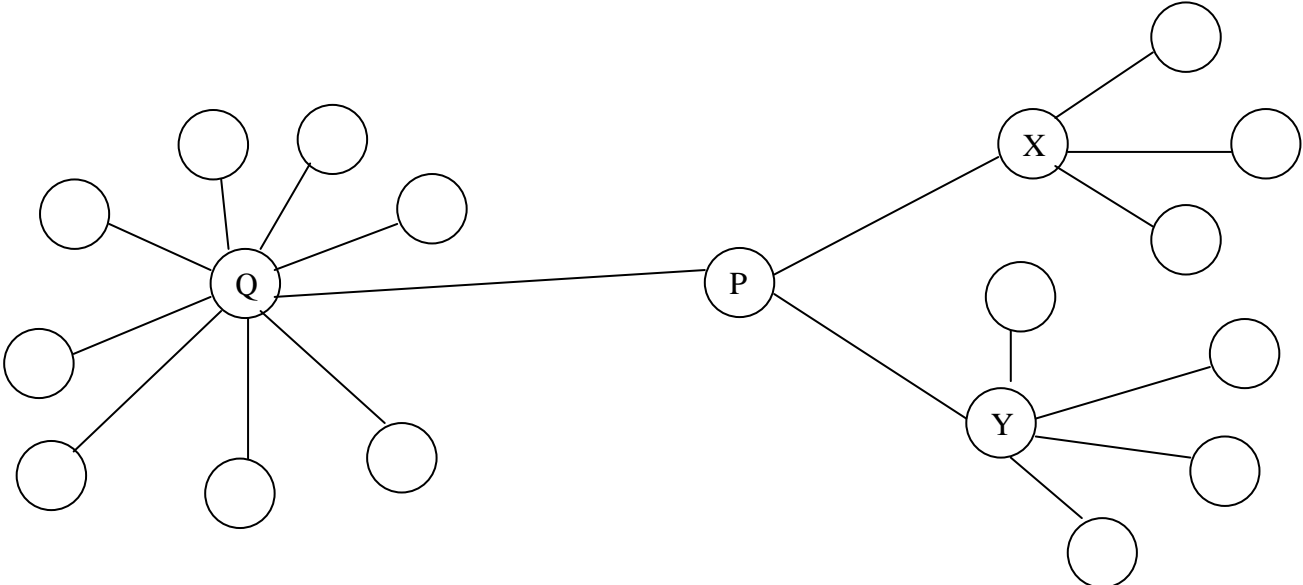
<sup>2</sup> Interestingly, Paul Erdos, who was one of the mathematicians who first developed random network theory, is also one of the hubs of twentieth century mathematics, with around 1400 papers and 500 co-authors. Many mathematics academics measure their *Erdos-number* – the number of co-authorship links between themselves and Erdos.

flow across the network even after severe degradation.<sup>3</sup> Broken links are not normally a problem in a multiply-connected network because other routes can usually be found. Failure of nodes can be more problematic. However, if the failure occurs randomly (the normal case for accidental or ‘natural’ causes), since most nodes only have a few links, the network will hardly be affected if they get knocked out. Even if one hub goes, the other hubs will be interconnected, so the information still flows. However, if the degradation is deliberate – the network is under attack – then it is the hubs that will be targeted. For a virus-type attack where each infected node passes attacks along all its links (especially towards the other hubs) the network can become **corrupted** very rapidly.

Having established that it is the hubs that are important for the functioning of a scale-free network, we need to consider how to measure the *relative importance* between different hubs – that is, we need to compare how **central** they are to the network as a whole (M. E. J. Newman, 2008). The most obvious measure of centrality is simply the number of links possessed by each node. This is known as the **degree** of the node. The hub with the greatest number of links has the highest degree and therefore would be considered the most central node. This is a pretty good measure for many scale-free networks. In particular it seems to be a good measure of centrality for the World Wide Web. The hubs that dominate the Web are the large search engines like Google and Yahoo. Since Google lists search hits primarily by degree (i.e. by the number of sites linked to each hit site in descending order) and the searcher tends to use the first site encountered that meets his or her requirement, Google tends to attract and promote other hubs, in order of popularity. This means that it is always worth going to Google first because it tends to find you the best’ (most popular) site for your query. This has given rise to concerns that Google may one day become so central that it will regulate all influential information transfer across the Internet. This phenomenon is known as ‘Googlearchy’ (M. Hindman et al, 2003).

There are other networks that are scale-free but where there are other nodes that may be more central than those of high degree. Consider the network depicted in 7th Figure. Although Q and Y and X have higher degree, the node P is more important (central) in the network because it joins a number of sub-networks that would otherwise be separated. This is rather an extreme case of a type of network whose particular shape means that one or a few nodes play an important linking role, even though their degree is not particularly high.

7th Figure: A betweenness network example



<sup>3</sup> In fact, one of the design goals of the original Internet (Arpanet) was to create a communications system for the USA that could continue to function even after a massive failure caused by, say, a nuclear attack on the country.

A **path** in a network is a route from one node to another, possibly passing through other nodes. The **geodesic path** between two nodes is the shortest path (the least number of links) between the nodes. Therefore, to measure how important (*central*) any given node might be, we define its **betweenness** as the fraction of all geodesic paths that pass through that node. For example, the network in Figure 7 has 162 geodesic paths and the node P lies on 110 of them, giving a betweenness of  $\sim 0.68$ . The node Q lies on 88 such paths so its betweenness is  $\sim 0.54$ . Betweenness gives a rough measure of the amount of information that flows through the node compared with the total information flow through the whole network. So a betweenness greater than 0.5 marks a very special node indeed (and one unlikely to be found in anything but rather small or unusually-shaped networks).

# Applying Network Analysis

---

One measure of success for a theory lies in its power to predict how events are likely to unfold. If scale-free networks realistically model the virtual society of the World Wide Web, then can we use network analysis to anticipate the way people's behaviour is influenced by the information flow in the Web? This is a question of great importance to the marketing departments of any large retail company, to the entertainment industry and all politicians undergoing elections. At the Massachusetts Institute of Technology, Peter Gloor has produced network analysis software called CONDOR that uses the betweenness of interconnected websites to make these kinds of predictions. For example, CONDOR was given the titles of all the popular films made in 2006, and its analysis generated a shortlist of the ten titles it identified as being the most impressive. Of these, five won Oscars and another four were nominated for Oscars (Jason Palmer, 2008). This is a good hit rate which many expert commentators would envy – how does it work?

CONDOR uses Google to do all the work. First, it finds the top ten sites that reference a particular film title. Then it finds the top ten sites that link to each of these sites. Then it constructs a network of all the mentioned sites, some of which may have been repeated and some of which may not even mention the original film title. CONDOR then assumes that these are the main hubs for that film which is reasonable because that is how Google works. Next CONDOR calculates the betweenness of all the sites in the constructed network and gives a score to the film based on the average betweenness value found. If a film is very popular – that is, a lot of sites are talking about it – then the most influential sites that discuss it are likely to be tightly connected and have high average betweennesses. So CONDOR's top ten titles would reflect this.

On the other hand, many websites may just consist of cinema listings or other routine marketing material (and not be concerned with what people are actually *discussing*), so for many cases this measure might be crude or inaccurate. By discriminating between the *type* of website (for example, blogs, discussion forums, general information sites) Gloor thinks he can refine his predictions. In particular he gives more weight (85%) to the chat forums used by “the most interested people talking to each other”. With this technique he was able to predict some stock market movements several days ahead with about 80% accuracy. Subsequently, some Italian researchers used CONDOR to rank candidates in a political election, using their popularity across various websites (Francesca Grippa & Pasquale Del Vecchio, 2008). When they compared the results gathered from google.com and technorati.com with the post-election results, they found no correlation. However, when they used Google Blog for the analysis, the correlation was positive. This reinforces Gloor's assertion about *interested people*.

By understanding the way people link up with one another on the Word Wide Web, it may be possible to make predictions about outcomes that depend on public opinion. Given the range and number of organizations that are interested in these predictions, social network analysis, applied to the virtual social networks that arise on the Web, may be a more acceptable way of delivering these results than the intrusive, personal data-capture and data-mining techniques that are currently employed by many multinational companies.

# Conclusion

---

This paper has suggested that the collection of websites that people access on the Internet correspond to the kinds of social networks that people join in the real world. By applying the techniques used to model social networks to this virtual world, it seems that real results may be attainable.

# Bibliography

---

Albert-Laszlo Barabasi – Eric Bonabeau (2003): Scale-free Networks (in: *Scientific American*, 288, 60-69, 2003)

*British and US Military Ranks* (<http://homepages.shu.ac.uk/~acsdry/ranks.htm>, accessed 20 May 2008)

Francesca Grippa – Pasquale Del Vecchio (2008): Take me to your Leader: Predicting Political Leadership using Social Network Metrics (in: *Proceedings of SUNBELT*, 28 Jan 2008)

M. Hindman – K. Tsioutsoulouklis – J. Johnson <sup>3</sup>Googlearchy<sup>2</sup>: How a Few Heavily-Linked Sites Dominate Politics on the Web ([www.cs.princeton.edu/~kt/mpsa03.pdf](http://www.cs.princeton.edu/~kt/mpsa03.pdf), accessed 12 April 2008)

Jason Palmer (2008): If you are connected you are a winner (in: *New Scientist* 2642, 07 February 2008)

M. E. J. Newman (2008): Mathematics of networks (in: L. E. Blume – S. N. Durlauf (eds.): *The New Palgrave Encyclopedia of Economics*, Palgrave Macmillan, Basingstoke)

*Spices and herbs go way back* (<http://www.home-herb-garden.com/herbandspiceroute.html>, accessed 20 April 2008)

*The Silk Road: Linking Europe and Asia Through Trade* (<http://library.thinkquest.org/13406/sr/>, accessed 20 April 2008)

Y. Chang – H. Makatsoris: *Supply Chain Modeling Using Simulation* (<http://ducati.doc.ntu.ac.uk/uksim/journal/Vol-2/No-1/>, accessed 20 May 2008)