Exystence
European Complex Systems Network of Excellence

COMPLEXITY WORKSHOP
HP Grenoble, Auditorium
November 28th 2005

Information flow in social networks

BERNARDO HUBERMAN, Senior HP Fellow and Director of the Information Dynamics Lab at Hewlett Packard Laboratories. Consulting Professor in the Department of Applied Physics at Stamford University.

Introduction

The flow of information in business organisations affects productivity and innovation because it determines the speed by which individuals can act and plan future activities. However, it may take place within social networks whose nature and existence is sometimes difficult to identify, because they are often different from what we might infer from the formal structure of a group or organisation.

E-mail, as the predominant means of communication, offers a unique opportunity to observe the flow of information along both formal and informal channels. It is a good medium for social network research, providing plentiful data in electronic form and allowing the discovery of shared interests and relationships which were not previously known. The tendency of individuals to associate according to common interests influences the way information spreads throughout a social group and, whilst personal privacy policies rule out direct associations between individuals and particular e-mail messages, analysis of links and word content can indicate collaboration and knowledge exchange. A number of tools have now been used to extract and aggregate the data enabling conclusions to be made about 'small world' dynamics, 'collaborative tagging' and 'viral marketing'.

Informal networks

Informal networks, or so-called 'communities of practice', naturally grow and coalesce within and outside organisations. Any institution that provides an efficient communication system eventually becomes threaded by communities of people who have similar goals and a shared understanding of their activities. These coexist with the formal structure of an organisation and serve many purposes such as the resolving of conflicting goals or problems and furthering the personal interests of members. They also provide effective means of individual learning and, with incentives, can enhance the productivity of the business organisation.

Identifying communities

Classical practice in discovering informal networks has been to gather data from interviews, surveys and other fieldwork, to discover links and communities by further analysis. Such methods are accurate, but time-consuming and labour-intensive. Computer analysis of a set of one million e-mail messages, collected over two months in the HP laboratories at Palo Alto, allowed researchers J. R. Tyler, D. M. Wilkinson, and B. A. Huberman [(1)] to construct a network of correspondence among some 400 people and enabled the discovery of communities by partitioning the network. The only information used from the e-mails were the names of the sender and the receiver, enabling the processing of a large number of e-mails while minimising privacy concerns. Qualitative evaluation of the data, consisting of face-to-face interviews, then validated the presence of such communities and provided interesting perspectives on those identified.

The method involves constructing graphs in which vertices represent individuals and lines or 'edges' between them represent a threshold number of e-mails sent. Communities are subsets of related vertices where many edges connect individuals of the same subset but there are few edges between the subsets. We can illustrate this discovery process with the following diagrams:

Partitioning of the graph is based on a principle of 'betweenness centrality', first proposed by Freeman [2] and defined by the number of shortest paths used in passing from one individual in a community to one individual in another. Thus the edge AB in figure 1. has high betweenness because all the paths between any circle and a square must pass through it. Its removal would mean that the squares and circles would split into separate communities. The repeated use of an algorithm identifies community edges of large betweenness successively removing them to resolve the graph into separated communities.

Identifying leadership roles
In addition to identifying the formal and informal communities of practice it is also possible to draw inferences about leadership in communities about which little is known. Individuals tend to organise both formally and informally into groups based on their common activities and interests.

Information flow in informal networks
The flow of information through informal networks is superficially similar to the way an infectious agent is transmitted among individuals; the pattern of contacts determining how far a disease spreads. There are however differences between information flows and the spread of viruses. Viruses infect any susceptible individual, whilst information is selective and passed by its host only to an individual the host thinks would be interested. Individuals with similar characteristics tend to associate with one another and individuals many steps removed in the network, on average, tend not to have much in common. Transmission probability depends upon the decay of similarity in individual characteristics. In particular the number of individuals that a given e-mail message reaches is very small in contrast with what would be expected on the basis of a virus epidemic model.

The analysis of a person-to-person recommendation network, consisting of 4 million individuals who made 16 million recommendations on half a million products showed that the propagation of recommendations and the cascade sizes could be explained using stochastic dynamic models .The growth of the recommendation network over time was assessed from the viewpoint of the sender and receiver of the recommendation in terms of purchases made. Whilst on average recommendations are not very effective at inducing purchases and do not spread very far, there are product and pricing categories for which viral marketing seems to be very effective.

'Small world' search
The observation that any two people in the world are most likely linked by a short chain of acquaintances, known as the 'small world' phenomenon, has been the focus of much research over the last 40 years. In an experiment in 2001 and 2002, 60,000 individuals were able to repeat the experiment using e-mail to form chains with just four links on average across different continents. A small world phenomenon is currently exploited by commercial networking services to help people network for both business and social purposes. Although many social ties are 'local', meaning that they are formed through a person's work or place of residence it has been shown that it takes only a few random links between people of different professions or location to create short paths in the social network.

Knowledge briefs
At HP consultants on one kind of project or another are encouraged to communicate with each other through something called 'knowledge briefs'. These are reports about different aspects which can be put on a common 'server' so that other consultants can see and use the information. By studying the access patterns of consultants looking for particular documents it is possible to find the common interests of individual consultants. Similar graphs to those previously shown can be used to show the number of users going to different documents where the colour of a vertex indicates a particular document and its size indicates the number of users accessing it. It is also possible to show that people accessing one document are also accessing a lot of other similar documents. The picture gives a lot of information about the kind of knowledge that a particular group is accessing. It is a very nice way of getting in touch with people who are doing similar work, but whose existence is not known.

PowerPoint analysis

Similar information on PowerPoint presentations can be used to find out what people are working on. Since all PowerPoint presentations are stored on a server it is possible to do a text analysis of all PowerPoint slides and use similarities in text to find similarities between people's interests. If, for example, an acronym such as RFID (radio frequency identification) is searched for in the PowerPoint data, a list is obtained of people ranked in terms of how many times the acronym occurs in the presentations that they have recently carried out. It is a very nice way of finding out which people have anything to do with a particular topic. A mouse click on a particular individual gives the network of the people that are linked to that person working on a particular topic and the shortest distance between yourself and any individual you might wish to contact.

Dealing with risk aversion

Most organizations, like most bureaucracies, tend to be risk averse, which means that people will not take on projects which have a low probability of success even though the end benefit may be high. This is because people are only rewarded for successful outcomes. One way of getting people to be more adventurous is through something called 'decision insurance'. It is a way of compensating people when their decision turns out to be wrong. It is, however, vulnerable to what is called 'moral hazard'. If people are compensated any decisions then there is no incentive to make right ones. However, by allowing the informal community closest to the person to monitor those decisions, moral hazard is effectively neutralised.

Collaborative tagging

The semantic Web is a grandiose vision of an information network connected by meaning. So far it has been more of a promise than a reality though there are local areas where it works fairly well. Collaborative tagging (also called folksonomy), depends on the idea that people tag anything that is of interest to them. It is a very fast growing trend. The journal *Nature* now has a system whereby scientists tag URLs that they think relevant to their particular interests. Everyone who then visits a particular URL can see the tags that other people have applied. It is essentially a social discovery process and one of the most famous is *Del.icio.us* which allows students from all over the United States to tag anything they like. For example they might tag a picture of a cat by calling it a 'cat' or a 'feline' or a 'pet'. Yahoo has a tagging system called *myweb* and the systems *CiteULike* and *Connotea* do the same thing for academic publications.

We might conclude that if people tag anything in any way they like we might simply end up with a chaotic collection of tags. However, this is not the case and a pattern emerges that can effectively be considered a semantic web. It is a bit like librarians asking everybody to tag the books in a library that they liked. Eventually the tags would start clustering around interesting groupings that the librarians would not have thought of. At HP Labs we used data that we got from *Del.icio.us* to determine when particular tagged books had reached a peak of popularity. Two examples are shown in the following diagrams:

The combined tags of many users also give rise to stable patterns in which the proportions of each tag become fixed:

The emergent classification system is now being used by the people that run Del.icio.us and Connotea.

How do we explain this phenomenon? (3) Suppose we have an urn containing two balls, one black and one white. We then blindfold ourselves and pick a ball. If it is a black ball we add another black ball and repeat the process. We might think that because we had initially picked a black ball the favoured probability would be to end up with all black balls. In fact the system settles to a constant ratio of black to white balls which is totally arbitrary and if we were to start the experiment again we would get another constant ratio of black to white balls bearing no relationship to what was obtained before. When people are influenced by just the number of other people's tags the classification is arbitrary, though extremely robust. If however, an item or a document means something to the tagger then very interesting semantic classifications emerge.

(1). J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "Email as spectroscopy: Automated discovery of community structure within organizations," in *Proceedings of the International Conference on Communities and Technologies*, (Netherlands), Kluwer Academic Publishers, 2003.

(2). L. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

(3). Golder, S. and Huberman, B. A.. (2006). “Usage Patterns of Collaborative Tagging Systems” Journal of Information Science, Vol. 32, pp203-213 (2006).