



# Graph & Network Analysis Cluster

## Research Overview

In the coming years, research in network data analysis will be transformed by access to large-scale data resources. This proposal presents a research programme in network analysis that addresses issues in internet services, fraud detection and bioinformatics.

A persistent theme in data analysis is the study of collections of entities and the links between these entities. This research has its origins in the work of Erdős and Rényi (1961) and Solomonoff and Rapoport (1951) and more recently in the work of Granovetter (1973, 1983) on “Weak Ties”. For a long time this strand of research on networks of actors was little more than an intellectual curiosity. It gave us the famous Erdős number game that is popular with mathematicians – the Erdős number for a researcher is the shortest number of co-author links that connect that researcher to Paul Erdős. At a more prosaic level it gave us “Six Degrees of Kevin Bacon”, a trivia game for connecting movie actors through ‘costarring’ relationships<sup>1</sup>.

In recent years work on social network analysis (SNA) has progressed to a new level. The field has now reached a stage where the models provide insights that are *actionable*. It is worth looking briefly at two recent studies that illustrate this. Perhaps the most impressive of these is the social network analysis that is part of the Framingham Heart Study (Christakis and Fowler, 2007, 2008). The SNA analysis in this study showed how influences from the social network of an individual influenced obesity and smoking patterns and that these influences only vanished beyond three degrees of separation. In a commentary on the obesity aspect of this study Barabási (2007) pointed out that social network characteristics are stronger predictors of obesity than some significant genetic markers. Considering the huge funding for research on prediction from genetic data, this is a remarkable outcome. Importantly, the authors of the study also show how insights derived from SNA provide strategies to reduce obesity and smoking.

In the field of human capital management and knowledge retention Davenport et al. (2006) have shown how a social network analysis of the communication structures in an organisation can identify employees that are key to the operation of the company. The authors review an analysis conducted in Delta Airlines and demonstrate how SNA can help frame strategies to reduce risks of knowledge loss due to the departure of key employees.

Closer to home, Idiró ([idiro.com](http://idiro.com)) who are partners in this research cluster have employed SNA in supporting a number of large European mobile phone operators in their customer retention and viral marketing activities for a number of years. They show that SNA helps predict *churn*, the loss of a customer to another operator. In the same way that a smoker in the Framingham study is more likely to quit smoking if his buddies quit, a mobile phone subscriber is more likely to switch networks if his friends have recently moved. The insights that SNA offers in these quite different areas depends on the ability to identify and model four key characteristics of networks;

- **Flow**, the way information and influence flows,

---

<sup>1</sup>[en.wikipedia.org/wiki/Six\\_Degrees\\_of\\_Kevin\\_Bacon](http://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon)

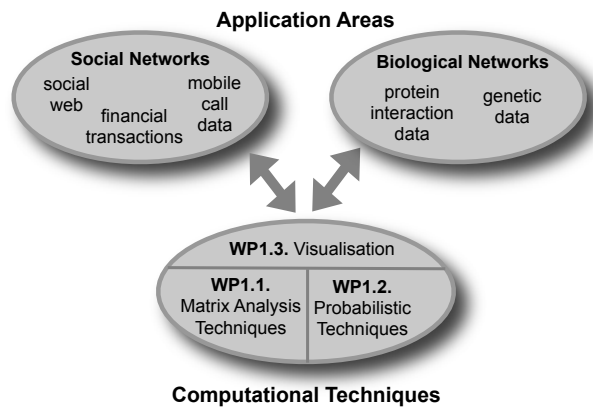


Figure 1: The overall research structure.

- **Communities**, the communities that exist,
- **Anomalous Structure**, the remarkable or anomalous structure that exists,
- **Centrality**, the nodes that are central and pivotal for communication.

The defining characteristic of social networks is the fact that at least some of the nodes in the network are human actors. In recent years there has been extensive research activity on the analysis of biological networks along these themes of flow, communities, anomalous structure and centrality (Greene et al., 2008; Jeong et al., 2001; Milo et al., 2002; Saul and Filkov, 2007). This analysis helps uncover functional structure, how signaling operates, and which genes or proteins are most crucial in a functional module. The transfer of techniques developed for social network analysis to biomedical data networks has proved very effective in recent research (Adamcsek et al., 2006; Jeong et al., 2001; Milo et al., 2002) This transfer of ideas is a two way street and there is considerable potential to use techniques that have been developed for motif discovery in biological networks for anomaly detection in social networks (Bryan et al., 2008). To foster this extensive cross-fertilisation, this cluster brings together both these key areas of network research.

**Research Structure:** This research cluster will conduct research on the themes of flow, communities, anomalous structure and centrality. The research programme has three research strands as shown in Figure 1. The central research strand will address the development of new techniques for the analysis and visualisation of network data. A special characteristic of this proposal is that it brings together experts on three aspects of network analysis from UCD and NUIG, i.e. experts on visualisation, matrix analysis techniques and probabilistic modelling. In addition to the work on computational techniques there will also be two applications research strands as shown in the figure.

The research programme is driven by a set of challenges identified by the three industrial partners; IBM Dublin Software lab, Norkom Technologies and Idiro Technologies and on the biological side by our collaborators in the UCD Conway Institute and in the Krogan Lab at the University of California-San Francisco. The industrial partners are key to this proposal as it is a characteristic of network analysis that the details of network connectivity influence the choice of analytic techniques. For this reason, access to real data is important to drive the research. A key strength of this proposal is that the industrial partners will provide information about the characteristics of real data. We

will also work on gathering (scraping) large social network datasets from the internet – DERI in NUI Galway have extensive experience in this area.

## Research Challenges

The three industrial partners participating in this research programme and their interests in network analysis are as follows:

- **Norkom Technologies:** ([norkom.com](http://norkom.com)) Norkom are an Irish company working on software and services for fraud and money laundering detection in financial services. They develop technology for detecting fraud in networks of transactions.
- **IBM Software Group:** IBM have located a significant part of their software development for social network analysis in Dublin and this will be a key academic collaboration for them.
- **Idiro Technologies:** ([idiro.com](http://idiro.com)) Idiró are an Irish company working on data analytics for the mobile telecoms market. Their core business is software for the analysis of mobile subscriber data.

In addition our collaborators, Prof. Des Higgins from the Conway Institute in UCD and Prof. Ger Cagney and Prof. Nevan Krogan from the University of California at San Francisco have identified research challenges in bioinformatics. In the remainder of this section we summarise this full set of research challenges and clarify the connection between these challenges and the network analysis research themes of flow, communities, anomalous structure and centrality.

**Financial Regulation & Fraud Detection:** Norkom Technologies work with financial institutions who need to carry out continuous transaction monitoring on all customers to ensure that any transactions indicative of suspicious activity are identified. Examples include; debit card and credit card fraud involving stolen or duplicated (skimmed) cards, mortgage application fraud to obtain a loan under false pretenses, and insurance claims fraud. Two key challenges relating to the analysis of such networks of data are:

1. Identifying anomalous payment structures. Many forms of financial crime e.g. taking over bank accounts through phishing attacks, involve routing unusual volumes of payments through specific sets of accounts within a payment network. In this instance, it is required to identify such unusual payment corridors i.e. specific nodes of the network through which an unusual volume of payment traffic flows.
2. Monitoring dynamic networks. It is also of interest to monitor a network as a dynamic entity. It is of specific interest to visualise networks which expand/contract over time, as well as providing automated methods to prune networks, or divide a large network into component parts according to defined criteria.

**Social Networking Software:** In this context social networking software covers not only the infrastructure software for social networking sites but also the email, instant messaging, blogging and the other components of the social web. Two challenges we will identify in this area are:

3. Discovering latent organisational structure from communication networks. Large organisations are interested in discovering latent company structure for reasons of human capital management. This helps expose vulnerabilities to resignations and retirements and to identify key individuals and bottlenecks in communication flow within the organisation.
4. Modelling and visualising how new ideas disseminate in communication mechanisms such as blogs.

**Mobile Network Data Analysis:** The subscriber and call data available to mobile network operators is a good example of the vast data resources that are part of the information age. Mobile operators are keen to analyse this data in order to address business problems such as churn prediction and marketing campaign management. Specific challenges we will address are:

5. Uncovering community structure in very large networks ( $10^6 - 10^9$  nodes) where communities can overlap and the evidence of community structure can come from more than one type of relation, e.g. call data and location information.
6. Visualising communities and connectivity in very large networks ( $> 10^6$  nodes).
7. Analysing and modelling diffusion of information, products and services in viral marketing campaigns.

**Biological Networks:** In modern biology large-scale experiments aimed at identifying the functions of the individual genes have generated a range of datasets of unprecedented complexity. The major challenges are integrating different data types obtained from different experimental approaches, defining the biological questions that need to be addressed, and matching the most appropriate computational approach to the question in hand. To focus our efforts onto real biological problems, we will use public data as well as data produced at the Krogan Lab at UCSF (physical and genetic protein interactions, drug sensitivity data), where high-throughput approaches are being applied to understand the activities of biological pathways under different conditions and across species. Two concrete challenges are as follows:

8. Integrating knowledge from physical and functional interaction networks with directed and undirected matrices of node properties,
9. Identifying and ranking local network features and clusters of biological and clinical interest.

# Bibliography

- B. Adamcsek, G. Palla, I. Farkas, I. Derenyi, and T. Vicsek. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, 2006.
- A. Barabási. Network Medicine—From Obesity to the” Diseasome”. *New England Journal of Medicine*, 357(4):404, 2007.
- K. Bryan, M. P. O’Mahony, and P. Cunningham. Unsupervised retrieval of attack profiles in collaborative recommender systems. ACM Recommender Systems (RecSys2008) (also available as UCD CSI Technical Report UCD-CSI-2008-03), 2008.
- N. Christakis and J. Fowler. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357(4):370, 2007.
- N. A. Christakis and J. H. Fowler. The Collective Dynamics of Smoking in a Large Social Network. *New England Journal of Medicine*, 358(21):2249–2258, 2008. doi: 10.1056/NEJMsa0706154.
- T. Davenport, R. Cross, and S. Parise. Strategies for preventing a knowledge-loss crisis. *MIT Sloan management review*, 47(4):31–38, 2006.
- P. Erdős and A. Rényi. On the evolution of random graphs. *Bulletin of the Institute of International Statistics*, 38:343–347, 1961.
- M. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6): 1360–1380, 1973.
- M. Granovetter. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1:201–233, 1983.
- D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics* (accepted for publication), 2008.
- H. Jeong, S. Mason, A. Barabasi, Z. Oltvai, et al. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks, 2002.
- Z. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604, 2007.
- R. Solomonoff and A. Rapoport. Connectivity of random nets. *Bulletin of Mathematical Biology*, 13(2):107–117, 1951.