

# Development and Application of a Directional Link Analysis Technique

黃勝偉、陳鴻文

E-mail: 9607621@mail.dyu.edu.tw

## ABSTRACT

Traditional techniques of social network analysis mainly focused on identify the central members in a network, but rarely explored the interaction and relationships between actors and sub-groups. A lot of information within a network is hence ignored. Therefore, a novel directional link analysis technique is proposed here. Instead, both dynamic and static relationships between actors and sub-groups will be investigated from the latent structure of a network. Enron e-mail corpus and indictment name list were used to illustrate and demonstrate the proposed techniques. First, the e-mail contact networks were classified by the topics of e-mail contents between July and Nov. 2001 which were regarded as the period of Enron fraud crisis. Following that, strongly connected components, communication bridges between components and other topological information of the e-mail networks were found by the proposed Directional Link Analysis (DLA). To identify the employee involving fraud within the node set found by DLA. The experimental results illustrated the effectiveness and efficiency of the proposed methods by the average detection rate and the average running time being 83.07% and 1.47 seconds, respectively. Therefore, DLA and DNSA are useful as novel tools of automatic link analysis because of their efficiency and objectiveness.

Keywords : 圖形理論(graph theory)、社會網絡分析(social network analysis)、方向性鏈結分析(directional link analysis)、安隆電子郵件分析(Enron corpus analysis)、弱鏈結(weak tie)

## Table of Contents

中文摘要.....	iii	英文摘要.....	v	誌謝辭.....	vii	內容目		
錄.....	viii	表目錄.....	x	圖目錄.....	xii	第一章 緒		
論.....	1	第一節 研究背景及動機.....	1	第二節 研究目的.....	3	第三節 研究範圍與		
限制.....	3	第四節 論文架構.....	4	第二章 文獻探討.....	6	第一節 鏈結分析的問		
題.....	6	第二節 圖形理論.....	13	第三節 社會網絡分析.....	17	第四節 安隆公		
司.....	23	第三章 研究方法與設計.....	31	第一節 研究方法與架構.....	31	第二節 方向性		
鏈結分析 (Directional Link Analysis, DLA).....	37	第三節 網絡結構差異分析(Differentiation of Network Structure Analysis, DNSA).....	41	第四章 實驗與結果評估.....	50	第一節 開發工具與設計環境.....	50	第二節 實驗資料來
源.....	50	第三節 實驗結果分析.....	54	第四節 結果評估比較.....	61	第五章 結論與未來		
研究方向.....	65	第一節 結論.....	66	第二節 後續研究發展建議.....	67	參考文		
獻.....	68	附錄A.....	81					

## REFERENCES

- Buchanan, M.(2003), 連結:混沌、複雜之後, 最具開創性的小世界理論, (胡守仁譯), 天下文化, (原文於2003年出版)。楊境榮(2005), 自動資料鏈結分析技術之開發與應用, 大葉大學資訊管理研究所未出版碩士論文。熊瑞梅(1995), 社會網絡的資料蒐集、測量及分析, 收於章英華、傅仰止、瞿海源編, 社會調查與分析:社會科學研究方法檢討與前瞻(pp. 313-356), 台北:中央研究院民族學研究所。二、英文部份
- Adriaans, P., & Zantinge, D. (1999). Data Mining. Addison-Wesley
- Bekkerman, R., McCallum, A. & Huang, G.. (2004). Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora. CIIR Technical Report IR-418, Available: <http://www.cs.umass.edu/~ghuang/foldering-tr05.pdf>.
- Berry, M. J. A., & Linoff, G.. (1997). Data Mining Techniques: For Marketing Sale and Customer Support. John Wiley & Sons.
- Berry, M. W. & Browne M. (2005). Email Surveillance Using Nonnegative Matrix Factorization. Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005. 45 – 54.
- Borgatti, S. P. (2004). The Key Player Problem. In: Dynamic So-cial Network Modeling and Analysis, Breiger, R., Carley, K.M., & Pattison, P., (Eds). National Academies Press, 241-252.
- Brass, D. J., & Burkhardt, M.E. (1992). Centrality and Power in Organizations. In: Networks and Organizations, Nohria, N., & Eccles, R.G., (Eds). Boston:Harvard Business School Press, 191-215.
- Burt, R. S. (1980). Models of Network Structure. Annual Review of Sociology, 6, 79-141.
- Burt, R. S. (1992). Structural Holes:The Social Structure of Competition. Harvard University Press, 45-49.
- Chapanond, A., Krishnamoorthy, M. S. & Yener, B. (2005). Graph Theoretic and Spectral Analysis of Enron Email Data. Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005. 15 – 22.
- Cohen, W. W. (2nd

ed.), CALO, CMU.

[Online], Available: <http://www-2.cs.cmu.edu/~enron/>. Diesner, J., & Carley, K. M. (2005). Exploration of Communication Networks from the Enron Email Corpus. *Proceedings of Workshop on Link Analysis*, 124-143. Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication Networks from the Enron Email Corpus " It's Always About the People. Enron is no Different ". *Computational & Mathematical Organization Theory*, 11(3), 201-228. Duan, Y., Wang, J., Kam, M. & Canny, J. (2002). A Secure Online Algorithm for Link Analysis on Weighted Graph. *Proceedings of Workshop on Link Analysis, Counter-terrorism and Security, SIAM International Conference on Data Mining 2005*. 71 – 81. Emirbayer, M., & Goodwin, J. (1994). Network Analysis, Culture, and the Problem of Agency. *American Journal of Sociology*, 99 (6), 1411-1454. Enron Dataset, Available: <http://www.isi.edu/~adibi/Enron/Enron.htm>, [2006, July 14]. Freeman, L. C. (1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks*, 1, 215-239. Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99, 7821-7826. Granovetter, M. S. (1973). The strength of weak ties. *Journal of American on Sociology*, 78, 1360-1380. Hanneman, R. A. (2001). A. Introduction to social network methods. California University Press, 87-105. Hansen, M. T. (1999). The search-transfer problem: the role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, 44, 82-111. Huberman, B. A., & Hogg, T. (1995). Communities of Practice: Performance and Evolution. *Journal of Computational and Mathematical Organization Theory*, 1(1), 73-92. Holme, P., Huss, M., & Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Journal of the Bioinformatics*, 19, 532-538. Johnson, M. P.; & Milardo, R. M. (1984). Network Interference in Pair Relationships: A Social Psychological Recasting of Slater's Theory of Social Regression Source. *Journal of Marriage and the Family*, 46(4), 893-899. Keila, P. S. & Skillicorn, D. B. (2005). Structure in the Enron Email Dataset. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*, 55 – 64. Klimt, B., & Yang, Y. (2004). The Enron Corpus: A New Dataset for Email Classification Research. In *Proceedings of ECML '04, 15th European Conference on Machine Learning*, 217-226. Klimt, B. & Yang, Y. (2004, a). Introducing the Enron Corpus. *First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, Available: <http://www.ceas.cc/papers-2004/168.pdf>. Klimt, B. & Yang, Y. (2004, b). The Enron Corpus: A New Data-set for Email Classification Research. *European Conference on Machine Learning*, Pisa, Italy. Knoke, D., & Burt, R. S. (1983). Prominence. In: *Applied Network Analysis: A Methodological Introduction*, Burt, R.S., & Minor, M.J., (Eds), 195-222. Knoke, D., & Kuklinski, J. H. (1982). *Network Analysis*. Sage, Beverly Hills. Knoke, D., & Kuklinski, J. H. (1991). *Network Analysis: Basic Concepts*. In: Thompson, G., Frances, J. and Levacic, R. (Eds). Sage, London, 173-182. Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. In N. Nohria, & Eccles, R. (Ed.), *Networks and organizations: Structure, form, and action*. Harvard Business School Press, 216-239. Krebs, V. E. (2002). *Unclouging Terrorist Networks*. *First Monday*, 7(4), 549-560. Laumann, E. O., Galaskiewicz, J., & Marsden, P.V. (1978). Community structure as interorganizational linkages. *Annual Review of Sociology*, 4, 455-484. Lazarevic, A., Ertöz, L., Ozgur, A., Srivastava, J., Kumar, V. A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection. *Proceedings of the 3rd SIAM Conference on Data Mining, 2003*, Available: [http://www.siam.org/meetings/sdm03/proceedings/sdm03\\_03.pdf](http://www.siam.org/meetings/sdm03/proceedings/sdm03_03.pdf) Luo, J. D., Yen, G. L., & Hui, W. S. (2003). Social Network Structure and Performance of Knowledge Team: A Case Study in the Chinese Cultural Settings. *Proceedings of Academy on Management*. Marsden, P. (1990). *Network Data and Measurement*. *Annual Review of Sociology*, 16, 435-463. McAndrew, D. (1999). The structural analysis of criminal networks. In: *The Social Psychology of Crime: Groups, Teams, and Networks*. Canter D., & Alison, L., (Eds). Dartmouth Publishing, Aldershot, UK, 53-94. McCallum, A., Corrada-Emanuel, A., & Wang, X. (2005a). Topic and role discovery in social networks. *Proceedings of the Nineteenth International Joint Conference*, 14, 786-791. McCallum, A., Corrada-Emanuel, A., & Wang, X. (2005b). The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks, with Application to Enron and Academic Email. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. 33-44. Milgram, S. (1967). The small world problem. *Psychology Today*, 2, 60-67. MIP. (2002). *Fortune Magazine's List of 10 Corporate Sins*. Mok, E. (2004). Enron Email Corpus: Mapping names to email addresses and doing network analysis. *Applied Natural Language Processing*. SIMS 290-2: Fall 2004, Prof. Marti Hearst, Assignment 4. Newman, M. E. J., & Girvan, M. (2003). Finding and evaluating community structure in networks. Michigan University press. Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. Michigan University press. Ouchi, W. G.. (1980). Markets, Bureaucracies, and Clans. *Administrative Science Quarterly*, 25, 129-141. Priebe, C. E., Conroy, J. M., Marchette, D. J. & Park Y. (2005). Scan Statistics on Enron Graphs. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. 23 – 32. Scott, J. (2000). *Social Network Analysis: A Handbook*. Sage, London. Scott, J. (2002). *Social networks: critical concepts in sociology*. Routledge, New York. Scott, W. R. (1992). *Organizations: Rational, Natural, and Open Systems*. Prentice-Hall. Seidam, S. B. (1983). Network Structure and Minimum Degree. *Social Networks*, 5, 269-287. Shetty, J., & Adibi, J. (2005a). Discovering Important Nodes through Graph Entropy: The Case of Enron Email Database. In *Proceedings of ACM SIGKDD LinkKDD*, 74-81. Shetty, J. and Adibi, J. (n.d., b), Ex employee status report. Retrieved November 4, 2004, from [http://www.isi.edu/~adibi/Enron/Enron\\_Employee\\_Status.xls](http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls). Shetty, J. and J. Adibi (n.d., c), The Enron Dataset Database Schema and Brief Statistical Report, Retrieved November 4, 2004, from [http://www.isi.edu/~adibi/Enron/Enron\\_Dataset\\_Report.pdf](http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf). Shetty, J. and J. Adibi (n.d., d), The Enron Dataset Mysql dump file, Retrieved November 4, 2004, from <ftp://ftp.isi.edu/sims/philpot/data/enron-mysqldump.sql.gz>. Sparrow, M. K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks*, 13, 251-274. SRI International's Artificial Intelligence Center, Cognitive Agent that Learns and Organizes Project, April 2006, [Online], Available: <http://www.ai.sri.com/people/gervasio>. Tichy, N. M., Tushman, M. L., & Fombrun, C.

(1979). Social network analysis for organizations. *Academy of Management Review*, 4(4), 507-519. Tyler, J. R., Wilkinson, D. M., & Huberman, B. A. (2005). Email as spectroscopy: Automated discovery of community structure within organizations. *Journal of the Information Society*, 21(2), 143-153. Wasserman, S., & Faust, K. (1997). *Social network analysis: Methods and application*. Cambridge University Press. Watts, D. J., & Strogatz, S. H. (1998). Collective Dynamics of Small-World. *Networks*, 1(393), 440-442. Wilkinson, D., & Huberman, H. (2002). A Method for Finding Communities of Related Genes. submitted for publication, <http://www.hpl.hp.com/shl/papers/communities/index.html>. Westphal, C., & Blaxton, T. (1998). *Data Mining Solutions*. John Wiley & Sons. Weiss, M. A. (1993). *Data Structures and Algorithm Analysis in C*. Benjamin Cummings.