Information and Productivity: Tools for Analyzing the Effects of Network Structure on Output

Charles King III Harvard Business School Morgan 187 Boston, MA 02163 (617) 495-1008	Marshall Van Alstyne University of Michigan 550 East Univ Ave. / 304 West Hall Ann Arbor, MI 48109 (734) 647-8028

Acknowledgments – This work has been generously funded by Intel and by NSF Career Award 9876233.

Introduction

Almost every major consulting firm offers a variation on IT benchmarking or knowledge management to help companies improve their resource use and gauge the efficacy of their IT investments. Despite the perceived importance that such ubiquitous offerings imply, general evidence for any productivity enhancements did not appear until 1996 (Brynjolfsson & Hitt). The continued absence of productivity evidence in the face of high investment was even characterized as a "productivity paradox" (Strassman 1990, Berndt & Morrison 1991, Brynjolfsson 1993, Loveman 1994, Landauer 1995).

The long term purpose of this study is to gather and test data at multiple business units to uncover information and network management practices that contribute to productivity at statistically significant levels. The technological impediments to this study are formidable and this talk reports on the first stage development of tools to observe information traffic flows across organizational networks. Ultimately, we will be using both hard technology, including physical hardware and software to measure packets, and soft technology, including surveys of user perceptions.

This abstract briefly describes a few of the hypotheses we will be examining, the technological barriers to instrumenting this study, and possible solutions to these barriers. We then present a few examples of network measures that have been successfully implemented.

Initial hypotheses will concern, for example, do balkanized and decentralized systems reduce output? Does universal intra-organizational access matter? Do incentives that promote information sharing result in more internal data capture? Does external data capture complement internal data capture? Does information overload exist and if so does it affect network boundaries?

Using network monitoring technology, we observe all packet traffic moving across a business unit's physical network. These packets are reassembled into message streams – allowing the tracking of e-mail, HTTP, video and other information streams – then mapped onto the social network of members in the organization. This technology was orginally developed by by a security firm to reassemble traffic histories following a network break-in.

A major advantage of this approach is the accuracy of data from direct observation. We are collecting data by monitoring actual information flows, eliminating reporting bias. Moreover, accounting statements, the traditional source of productivity data, typically provide less direct evidence of how organizations use information than system level measures. These data will be supplemented by data from financial statements and interviews that should provide a rich context to explain how information systems are perceived.

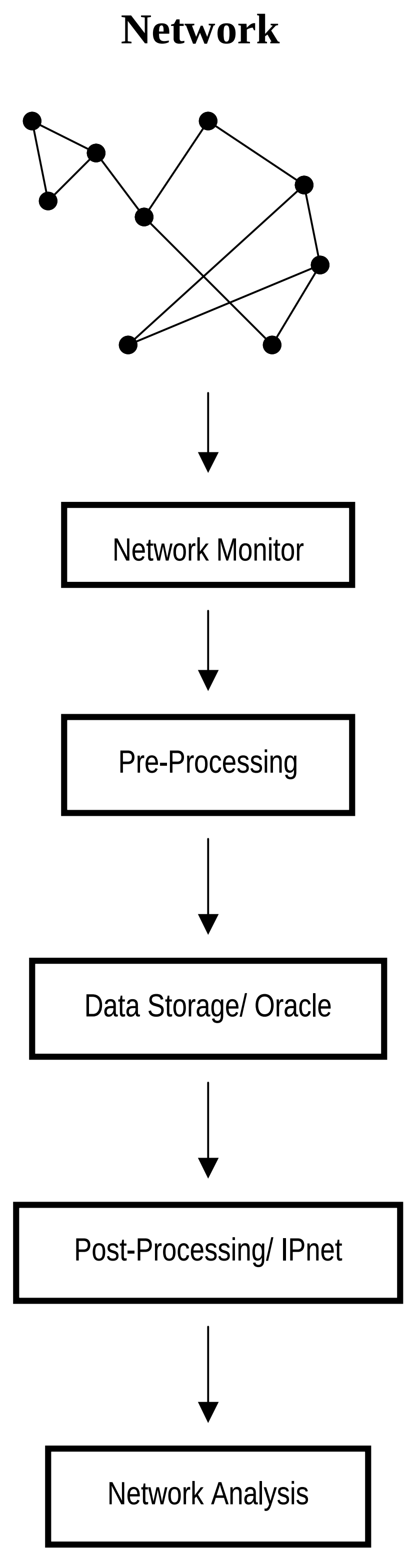


Figure 1 depicts the general steps of the data gathering process. The network monitors observe the packet transfers directly. Pre-processing discards unnecessary packets, such as a printer announcing its availability to the net, and filters overhead under strict real time requirements. The data are then formatted for storage in an Oracle database. This step also adds tables that give us convenient data for further analysis, such as an IP host table that lists the IP address, host name, and department or user information. Post-processing then corrects for such anomalies as one person using multiple machines. It also reformats data as necessary to input it into other tools. The data can then be fed into network plotting tools such as UCINet and KrackPlot to perform social network analysis

Issues

This study faces several major obstacles that are social as well as technical.

Data Capture: With the rise of 100Mbit and greater network speeds, the recording instruments themselves can become the bottleneck for data gathering. The massive amount of data also introduces severe data storage and computational problems as a few days worth of traffic can generate gigabytes of data. Finally, the instrumentation should not introduce overhead that interferes with performance. To gather high-speed data, we use separate recording stations capable of buffering significant traffic before writing to disk. Additionally, we only need to capture data sufficient to preserve the statistical distribution of traffic, not to recreate entire histories. This allows us to drop packets in a prescribed fashion although algorithmic computation on very large networks is still proving difficult. Since monitoring stations are independent, their computation is not experienced by existing hardware and they can fail without introducing delays in user traffic.

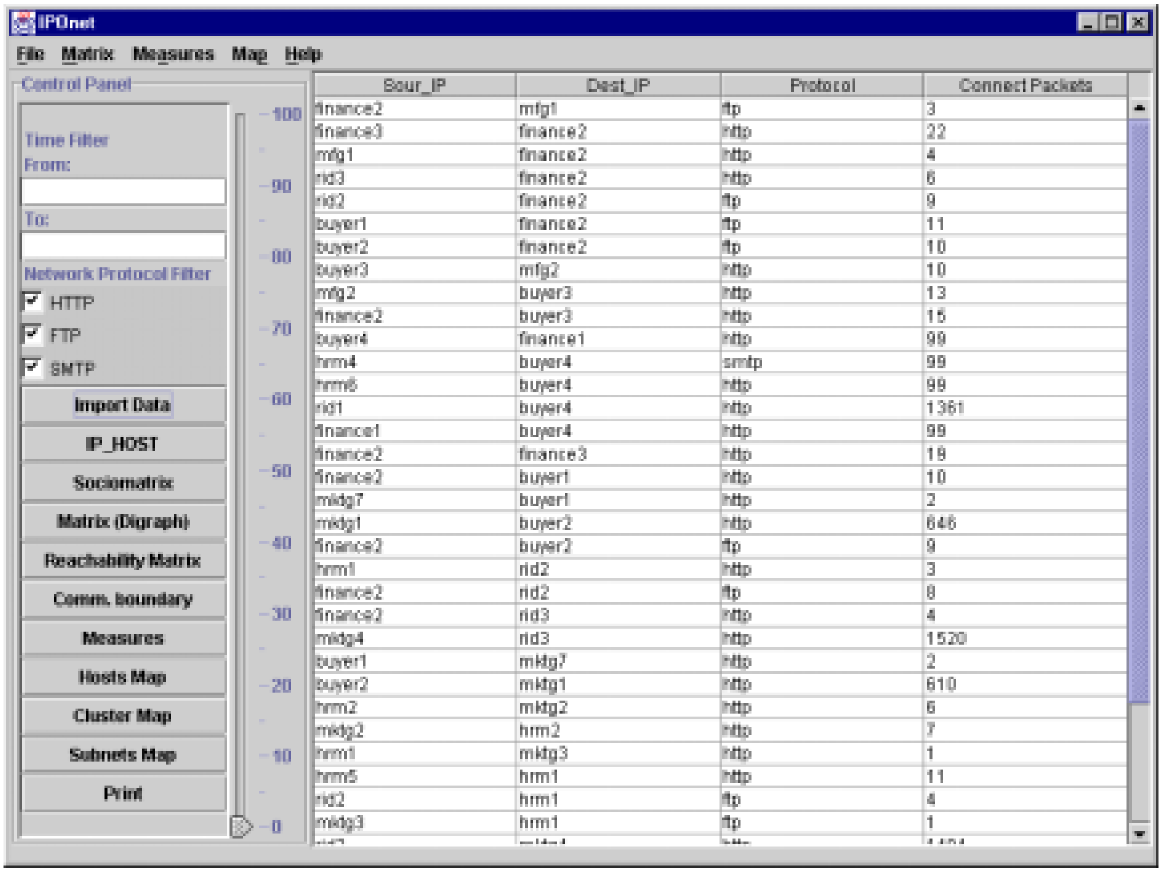
Packet Switched Networks: As single packets in a message stream need not follow a single route, data gathering becomes difficult without a single point for making observations. Packet switched networks are much more difficult circuit switched networks. We resolve this by observing traffic from multiple locations and by using repeaters to direct traffic to network monitors. We need to use enough network monitors or repeaters to give a complete picture of network flows.

Logical versus Physical Networks: Our primary interest concerns how individuals communicate and share information, the logical network, but the tools observe the machinery in use, the physical network. This introduces complications from routing and from mapping people to machines. E-mail traffic, for example, is typically routed to a central server then to the recipient. On the physical network, the server appears as a central node but, on the logical network, it is largely irrelevant. We use mappings of IP addresses and user accounts to bump up a level to the logical network. Similarly, if individuals use multiple machines, for example, a home and travel computer, then we also need a mapping to associate people and machines. These maps are typically provided by network administrators and user account logs.

Privacy and Intrusiveness: Our technology makes it possible to reconstruct individual message histories, raising privacy and auditing concerns. We handle this in three ways. First, we can keep only statistical samples of packet movement as distinct from entire histories. We need only preserve the statistical distribution of packet movement to eliminate sample bias and preserve the validity of regressions. This prevents us, however, from reconstructing actual sessions. Second, both IP addresses and individual names are mapped to pseudonymous values, maintained by an independent party. We cannot associate real people or machines with individual packets. Research requests for more information from any individual, can pass through this independent party without revealing these identities. Third, we encrypt our data so that, even if it were misappropriate by an ill-intentioned third party, misuse would be exceedingly difficult.

Preliminary Tools

A prototype is partially functional and is built on top of a network scripting language, Java, and Oracle. Compared to the speed of C or C++, the speed of Java is somewhat slower, but programming with Java gives us the flexibility to use the application remotely using browsers and different systems. The 2-D graphing features of Java 2 are also very helpful in recreating network visualizations.



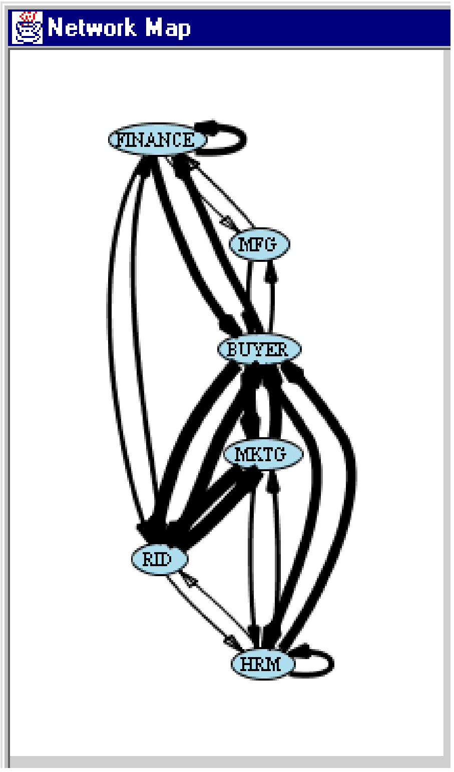
Sour_IP	Dest_IP	Protocol	Connect Packets
finance2	mfg1	ftp	3
finance3	finance2	http	22
mfg1	finance2	http	4
rid3	finance2	http	6
rid2	finance2	ftp	9
buyer1	finance2	ftp	11
buyer2	finance2	ftp	10
buyer3	mfg2	http	10
mfg2	buyer3	http	13
finance2	buyer3	http	15
buyer4	finance1	http	99
hrm4	buyer4	smtp	99
hrm6	buyer4	http	99
rid1	buyer4	http	1361
finance1	buyer4	http	99
finance2	finance3	http	19
finance2	buyer1	http	10
mktg7	buyer1	http	2
mktg1	buyer2	http	646
finance2	buyer2	ftp	9
hrm1	rid2	http	3
finance2	rid2	ftp	8
finance2	rid3	http	4
mktg4	rid3	http	1520
buyer1	mktg7	http	2
buyer2	mktg1	http	610
hrm2	mktg2	http	6
mktg2	hrm2	http	7
hrm1	mktg3	http	1
hrm5	hrm1	http	11
rid2	hrm1	ftp	4
mktg3	hrm1	ftp	1

The result of "Import Data"

These data were generated using anonymous traffic on a small volunteer site. IPnet currently works with networks of one hundred nodes. Additional trials on larger networks will indicate how performance scales when the number of nodes increases dramatically.

The next figure shows the network connections that emerge from the packet traffic. Line width indicates the frequency of communication

There are several types of results and extensions from this study. Confirmation or rejection of one or more hypotheses may help improve managerial practice. Since few results are firmly established, even failed hypotheses are likely to add value. By ruling out seemingly logical practices (for which no supporting evidence can be found), useful results could be obtained that indicate what not to try. On the other hand, affirmative results – indicating what managers should do – would likely be highly valuable. Either way, the study provides at least potential benefit.



The academic evidence for productivity enhancement is strong but recent. Given that information technology appears to be effective, there remains much to be understood about how and why this happens so that there appear to be numerous opportunities to develop the theories for why the various results obtain.

References

Berndt, E.R., & Morrison, C.J. (1991, December 9). Computers Aren't Pulling Their Weight. Computerworld, pp. 23-25.

Brynjolfsson, E. & Hitt, L. (1996). Paradox Lost? Firm-level Evidence of Returns to Information Systems Spending. Management Science, 42(4), 541-558.

Freeman, L. (2000). Visualizing Social Networks. Journal of Social Structure, 1(1).

Krackhardt, D., Bylthe, J. & McGrath, C. (1995). KrackPlot 3.0 User's Manual. Pittsburgh: Carnegie-Mellon University.

Landauer, T. (1995). The Trouble with Computers: Usefulness, Usability, and Productivity. Cambridge, MA: MIT Press.

Loveman, G. (1994). An Assessment of the Productivity Impact of Information Technologies. in T. Allen & M. Scott Morton (Eds.), Information Technology and the Corporation of the 90s, (pp. 84-110): Oxford University Press.

Sanil, A. Banks D., Carley, K. (1995) "Models for Evolving Fixed Node Networks: Model Fitting and Model Testing" Social Networks 17 pp 65-81.

Strassman, P.A. (1990). The Business Value of Computers. New Canaan, Conn: Information Economics Press.

Wasserman, S. & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

Zhiang, L. & Carley, K.M. (1997) Organizational Response: The Cost Performance Tradeoff. Management Science. 43(2), 217-234.