

## **Network Structure & Information Advantage: Structural Determinants of Access to Novel Information**

Sinan Aral (MIT) • Erik Brynjolfsson (MIT) • Marshall Van Alstyne (Boston University & MIT)  
*Workshop on Information Systems Economics, 2006*

### **Introduction**

Since the 1950s, economists and economic sociologists have consistently demonstrated a robust link between structural properties of individuals' and groups' networked relationships and measures of performance. However, the mechanisms driving this link, thought to be related to the value of the information flowing between connected actors, have thus far only been inferred, not empirically demonstrated.

Burt (1992) shows that individuals with structurally diverse networks (i.e. networks low in (a) redundancy, and (b) structural equivalence) are more successful (in terms of wages, promotion, job placement, and creativity). Aral, Brynjolfsson and Van Alstyne (2006) demonstrate that structural diversity is associated with higher levels of economic productivity for task based information workers. These studies infer that network diversity is associated with performance, in part because diverse contacts provide access to novel information and resources. Novel information is valuable due to its local scarcity. Actors with scarce, novel information in a given local network neighborhood may be better able to broker opportunities, use information as a commodity, or apply information to a problem that is intractable given local information – creating value and conferring advantages on those with more novel information. While theories of the value of information and empirical evidence on the relationship between network structure and performance exist, little theory, and almost no empirical evidence addresses the relationship between network structure and the nature of information distributed across a network. Establishing a link between these two literatures requires theory and evidence on the mechanisms by which social structure affects the distribution of novel information. Uncovering the mechanisms that drive the distribution of information across networks is critical to advancing our understanding of how social structure impacts performance.

To address this critical inference, we build a theoretical framework linking network structure to the distribution of novel information among actors, and test implications of our theory using empirical evidence from a longitudinal dataset containing the email communication patterns and message content of a group of information workers in a medium sized executive recruiting firm over a period of ten months. Our preliminary findings indicate that: (1) The *amount* of novel information flowing to an actor is *increasing* in the actor's *network size* and *network diversity*; (2) The *diversity* of information flowing to an actor is *increasing* in the actor's *network size* and *network diversity* controlling for the total amount of communication, while the marginal increase in information diversity is decreasing in network size. Part of the explanation for the decreasing marginal contribution of network size to information diversity is that: (3) In bounded networks, structural diversity is increasing in network size, but with diminishing marginal returns. As actors establish relationships with a finite set of alters, the probability that a marginal relationship will be non-redundant decreases as possible alters in the network are exhausted. We also find that (4) traditional human capital variables (e.g. age, gender, industry experience, education) have little effect on access to diverse information, highlighting the importance of network structure for information advantage. These results represent some of the first empirical evidence on the relationship between network structure and characteristics of the information content flowing to and from actors in a network.

### **Information, Communication and Productivity**

Understanding the role of information in productivity and performance is an important endeavor for economists, economic sociologists and information systems researchers. In the economics literature, the role of information and its distribution across actors has been a central concern for some time (e.g. Akerlof 1970, Marshak & Radner 1972, Arrow 1985). Previous research has also applied social network analysis to examine the performance of corporate R&D teams (Reagans and Zuckerman 2001), the output of new product development units (Hansen 1999, 2002), and the flow of information in R&D labs (Allen 1977) among other things. By mapping the structure of individuals', teams' or firms' direct and indirect ties, these studies empirically test the relationship between networked social structure and quantifiable dependant variables like productivity (Aral, Brynjolfsson and Van Alstyne 2006), project completion time (Aral, Brynjolfsson and Van Alstyne 2006, Hansen 2002), innovation output (patent applications accepted per unit time) (Ahuja 2000), as well as managers' job performance, mobility and well being (Burt, 1992; Podolny and Baron, 1997).

While these studies regard the value of communication content to be the driver of performance benefits from different structural patterns, none directly measure communication content. Building on our earlier work on the relationship between network structure and productivity (Aral, Brynjolfsson and Van Alstyne 2006), we examine the relationship between network diversity and information diversity and ask whether information diversity explains the relationship between networks and productivity.

### Models and Hypotheses

In a social network of  $n$  individuals represented as a directed graph  $G$  with  $n$  nodes, messages  $m$  flow to and from individuals. The number of contacts of individual  $i$ , with whom  $i$  exchanges at least one message, is referred to as the size  $S_i$  of  $i$ 's network. We define the structural constraint  $C_i$  (Burt 1992: 55)<sup>1</sup> of an actor's network as the degree to which an individual's contacts are connected to each other (a proxy for the redundancy of contacts), such that  $C_i = \sum_j \left( p_{ij} + \sum_q p_{iq} p_{qj} \right)^2$ ,  $q \neq i, j$ ; and the structural diversity  $D_i$  of an actor's network as  $1 - C_i$ . The total amount of  $i$ 's incoming communication is  $E_i^I$ , such that  $E_i^I = \sum_j m_{ji}$ , where  $m_{ji}$  represents a message sent from  $j$  to  $i$ . We represent the diversity of the information in a given set of messages  $m$  as  $\alpha_m$ , where  $0 \leq \alpha_m \leq 1$ , and the diversity of the information in a given actor  $i$ 's inbox as  $\alpha_i^I$ . The total amount of non-redundant information flowing to actor  $i$  is referred to as  $NRI_i^I = f(\alpha_i^I, E_i^I)$ .

As individuals communicate with more contacts, and as individuals' networks connect them to actors that are themselves unconnected, we expect the information they receive to be more diverse and for them to receive more total novel information.

*H1: Individuals' network size and structural network diversity are associated with greater information diversity and with access to more non-redundant information.*

However, in bounded networks, as individuals add contacts to their networks, the probability that an additional contact will have novel information is likely decreasing in the size of the network. Assuming there is at least some overlap between actors' information, the marginal contribution of novel information from an actors' contacts should decrease in the number of contacts. We therefore expect the marginal increase in information diversity is decreasing in size:

*H1b: The marginal increase in information diversity is decreasing in network size. (i.e.:  $\frac{\partial \alpha_i^I}{\partial S_i} \geq 0$ ;*

*$\frac{\partial^2 \alpha_i^I}{\partial^2 S_i} \leq 0$ ). Furthermore, if connections between actors predict greater information overlap, the*

decreasing marginal contribution of novel information from an additional contact should be even sharper. As actors establish relationships with a finite set of alters, the probability that a marginal relationship will be structurally non-redundant should decrease as possible alters in the network are exhausted. Therefore:

*H2: The marginal increase in individuals' structural network diversity is decreasing in network size.<sup>2</sup>*

(i.e.:  $\frac{\partial D_i}{\partial S_i} \geq 0$ ;  $\frac{\partial^2 D_i}{\partial^2 S_i} \geq 0$ ). Finally, it could be that as individuals gain experience, they collect

expertise across several domains, reflected in communications across multiple subjects or topics:

*H3a: Individuals' information diversity and access to non-redundant information are increasing in their age, education and industry experience.*

However, it could also be that as individuals gain experience, they specialize and focus their work and their communication on a limited number of topics. Thus, we offer a competing hypothesis:

*H3b: Individuals' information diversity and access to non-redundant information are decreasing in their age, education and experience.*

<sup>1</sup> Where  $p_{ij} + \sum_q p_{iq} p_{qj}$  measures the proportion of  $i$ 's network contacts that directly or indirectly involve  $j$  and  $C_i$  sums this across all of  $i$ 's contacts.

<sup>2</sup> Although we do not address density in this abstract, we also expect that the density of connections in a network will amplify this effect.

## Measuring Information Diversity

We measured the diversity of information in individuals' email inboxes and outboxes using a vector space model of topics present in email content. We represented each email as a feature vector of topic keyword frequencies, grouped emails by individuals' incoming and outgoing mail and then used a series of five independent measures to assess the variance or spread of the vectors in a given inbox or outbox in multidimensional topic space. For example, an email about pets might include two mentions of the word "dog," two of "cat," and three of "veterinarian;" while an email about econometrics might include three uses of the word "variance," two of "specification," and three of "heteroskedasticity." Emails about similar topics should contain similar language on average, and the vectors used to represent them should therefore be closer in multidimensional space, reducing the variance or spread between them.

Instead of imposing exogenous keywords on the topic space, we extracted topic keywords from our email corpus using a series of algorithms guided by two key principles – that keywords: (1) distinguish topics (i.e. have a high frequency variance across topics), and (2) represent topics (i.e. have high frequency and low frequency variance within topics). We used a k-means clustering algorithm embedded in IBM's eClassifier tool to create the initial clusters used to guide our keyword selection.

Using these keywords, we created diversity measures based on the variance of the cosine similarity of email vectors, the variance of document similarity measures such as Dice's coefficient, measures of the relative clustering of emails using eClassifier and these clustering measures enhanced by an information theoretic weighting of emails based on their 'information content.' As all metrics produced highly correlated measures of diversity ( $\sim \text{corr} = .98$ ), our statistical specifications use the cosine distance deviation from the mean vector to represent the diversity of information in an email inbox at a given time.

$$\alpha = \frac{1}{\text{number\_of\_documents}} \sum_{d \in \text{documents}} (\text{MeanDistCos}(d))^2 ; \text{ Where } \text{MeanDistCos}(d) = \text{CosDist}(d, M)$$

We validated our diversity measurement by creating a spectrum of high to low diversity clusters of documents from Wikipedia.org. For example, we created a minimum diversity cluster using a fixed number of documents from the same sub category of the Wikipedia topic hierarchy, and a maximum diversity cluster using the same number of documents chosen from different sub categories. As Wikipedia contains several layers of topic hierarchy, we constructed a series of document clusters ranging from low to high topic diversity.<sup>3</sup> We then used this document corpus to generate keywords and measure diversity using the methods described above. Our methods were very successful in characterizing the diversity of these document clusters and appropriately ranking them from low to high diversity.

## Data

Our data cover 10 months of complete email history at a mid-sized executive recruiting firm. The data were captured from the corporate mail server during two equal periods from October 1, 2002 to March 1, 2003 and from October 1, 2003 to March 1, 2004. Participants received \$100 in exchange for permitting use of their data, resulting in 87% coverage of recruiters eligible to participate and more than 125,000 email messages captured. Details of data collection can be found in Aral, Brynjolfsson & Van Alstyne (2006).<sup>4</sup>

## Statistical Specifications & Preliminary Findings

We first examined the relationship between network structure and the diversity of information flowing into actors' email inboxes.<sup>5</sup> We tested both pooled OLS specifications controlling for individual characteristics with standard errors clustered by individual, and fixed effects models on monthly panels of individuals' networks and information diversity. We focused specifically on the size and structural diversity of individuals' networks and controlled for temporal variation by month.

$$\alpha_i^l = \gamma_i + \beta_1 E_i^l + \beta_2 S_i + \beta_3 S_i^2 + \beta_3 D_i + \sum_j B_j \text{HumanCapital}_i + \sum_m B_m \text{Month} + \varepsilon_i$$

<sup>3</sup> To control for variance created by the total number of documents, we constructed clusters of varying sizes.

<sup>4</sup> *F*-tests comparing performance levels of those who opted out with those who remained did not show statistically significant differences. *F* (Sig): Rev02 2.295 (.136), Comp02 .837 (.365), Multi .386 (.538).

<sup>5</sup> We expect network structure to influence incoming information more than outgoing. In future work, we will examine differences between incoming and outgoing information.

Our results demonstrate that the *diversity* of information flowing to an actor is *increasing* in the actor's *network size* and *network diversity* controlling for the total amount of communication, while the marginal increase in information diversity is decreasing in network size (Models 1,2).

We then examined the relationship between network structure and the total amount of novel information flowing into actors' email inboxes. We again tested pooled OLS and fixed effect specifications of individuals' networks and information diversity using the following model:

$$NRI_i^I = \gamma_i + \beta_1 S_i + \beta_2 S_i^2 + \beta_3 D_i + \sum_j B_j HumanCapital_i + \sum_m B_m Month + \varepsilon_i$$

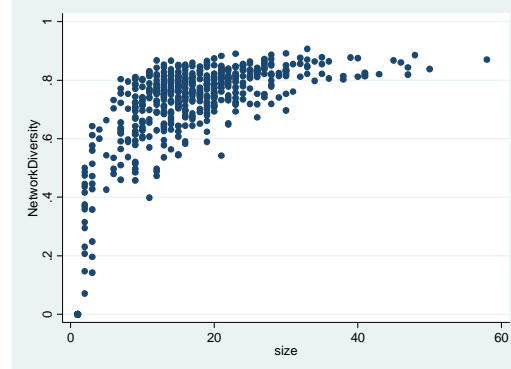
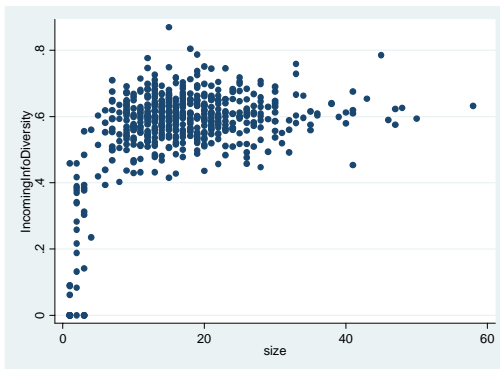
Our results demonstrate that the *amount* of novel information flowing to an actor is *increasing* in the actor's *network size* and *network diversity* (Models 3-6). Network diversity has a strong positive relationship with the total amount of novel information flowing into actors' inboxes, but is not significant when controlling for network size. The impact of size on total novel information dominates that of structural diversity because of the strong relationship between size and the total amount of incoming email, a critical driver of the total amount of novel information.

Finally, to explore the mechanisms driving the non-linear relationship between network size and information diversity, we tested our hypothesis that while structural diversity is increasing in size, there are diminishing marginal diversity returns to size in bounded networks. As actors establish relationships with a finite set of alters in a bounded network, the probability that a marginal relationship will be non-redundant should decrease as possible alters in the network are exhausted. If this is the case, we should see a non-linear positive relationship between network size and structural diversity in our data, such that the marginal increase in structural diversity is decreasing in size. To test this hypothesis, we specified the following model:

$$D_i = \gamma_i + \beta_1 S_i + \beta_2 S_i^2 + \sum_j B_j HumanCapital_i + \sum_m B_m Month + \varepsilon_i.$$

As our results demonstrate, there is a strong, non-linear, positive relationship between network size and structural diversity in our data, indicating that structural diversity is increasing in network size, but with diminishing marginal returns (Models 7, 8).

	<b>Model 1</b>	<b>Model 2</b>
<i>Dependent Variable:</i>	<b>Incoming Information Diversity</b>	<b>Incoming Information Diversity</b>
<i>Specification</i>	<i>Fixed Effects</i>	<i>OLS-c</i>
Total Incoming	-.0000	.0001
Email ( $E_i^I$ )	(.0001)	(.0001)
Size ( $S_i$ )	.0881*** (.0201)	.0695** (.0334)
Size Squared ( $S_i^2$ )	-.0623*** (.0159)	-.0666** (.0268)
Network Diversity ( $D_i$ )	.0568*** (.0066)	.0708*** (.0136)
Constant ( $\gamma_i$ )	.5688*** (.0122)	.6213*** (.1121)
Temporal Controls	Month	Month
Individual Controls	-	Gender, Age, Education, Industry Exp., Managerial Level
F-Value (d.f.)	20.56*** (12)	10.59*** (18)
R <sup>2</sup>	-	.48
Obs.	563	448



	<b>Model 3</b>	<b>Model 4</b>	<b>Model 5</b>	<b>Model 6</b>	<b>Model 7</b>	<b>Model 8</b>
<i>Dependent Variable:</i>	<b>Incoming Non-Redundant Information</b>	<b>Incoming Non-Redundant Information</b>	<b>Incoming Non-Redundant Information</b>	<b>Incoming Non-Redundant Information</b>	<b>Network Diversity</b>	<b>Network Diversity</b>
<i>Specification</i>	<i>Fixed Effects</i>	<i>OLS-c</i>	<i>Fixed Effects</i>	<i>OLS-c</i>	<i>Fixed Effects</i>	<i>OLS-c</i>
Size ( $S_i$ )		26.4743*** (4.3161)		42.5340*** (8.4625)	1.5851*** (.1127)	1.6282*** (.2093)
Size Squared ( $S_i^2$ )		-4.9533 (3.5161)		-19.1924** (9.4928)	-1.0381*** (.0982)	-1.0695*** (.1904)
Network Diversity ( $D_i^I$ )	7.4725*** (1.4069)	-.7578 (1.4660)	17.2689*** (2.9954)	-2.9080 (2.2770)		
Constant ( $\gamma_i$ )	38.6168*** (2.4110)	37.5306*** (2.0658)	99.4772** (34.9564)	121.7967*** (29.0196)	.0827 (.0639)	.6507 (.6303)
Temporal Controls	Month	Month	Month	Month	Month	Month
Individual Controls	-	Gender, Age, Education, Industry Exp., Managerial Level	-	Gender, Age, Education, Industry Exp., Managerial Level	-	Gender, Age, Education, Industry Exp., Managerial Level
F-Value (d.f.)	12.66*** (9)	30.64*** (11)	15.43*** (15)	13.40*** (17)	33.39*** (10)	15.58*** (16)
R <sup>2</sup>	-	-	.37	.56	-	.64
Obs.	563	563	448	448	563	448

### Conclusions & Next Steps

We present some of the first empirical evidence on the relationship between network structure and characteristics of the content of information flowing to and from actors in a network. As next steps we intend to assess the degree to which connections between actors sharpen the diminishing diversity returns to network size relative to that generated by random information overlap between actors; to test impacts of dimensions of information content and network structure on employees' productivity and effectiveness; and to extend analysis of network structure to a broader set of structural characteristics beyond size and diversity.

### References

- Ahuja, G. (2000). "Collaboration networks, structural holes and innovation: A longitudinal study." *Administrative Science Quarterly* 45: 425-455.
- Akerlof, G. (1970). The market for lemons: quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84 (3), 488-500.
- Allen, T. J. (1977). *Managing the flow of technology*. Cambridge, MA, MIT Press.
- Aral, S., Brynjolfsson, E., & Van Alstyne, M. (2006). "Information, Technology and Information Worker Productivity: Task Level Evidence." *Proceedings of the 27<sup>th</sup> Annual International Conference on Information Systems*, Milwaukee, Wisconsin.
- Arrow, K (1984). "Informational Structure of the Firm." *American Economic Review* 75(2): 303-307.
- Burt, R. (1992). *Structural holes*. Cambridge, MA, Harvard University Press.
- Hansen, M. (1999). "The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits." *Administrative Science Quarterly* 44(1): 82111.
- Hansen, M. (2002). "Knowledge networks: Explaining effective knowledge sharing in multiunit companies." *Organization Science* 13(3): 232-248.
- Marschak, J. & Radner, R (1972). *Economic Theory of Teams*, Yale University Press, New Haven, 1972.
- Podolny, J. and J. Baron (1997). "Resources and relationships: Social networks and mobility in the workplace." *American Sociological Review* 62(5): 673-693.
- Reagans, R. and E. Zuckerman (2001). "Networks, diversity, and productivity: The social capital of corporate R&D teams." *Organization Science* 12: 502-517.